UNIVERSIDADE DE CAXIAS DO SUL
ÁREA DO CONHECIMENTO DE CIÊNCIAS EXATAS E ENGENHARIAS

MATHEUS SAUTHIER

MODELOS PREDITIVOS PARA PREVISÃO DE CRISES FINANCEIRAS

CAXIAS DO SUL

2021

MATHEUS SAUTHIER

MODELOS PREDITIVOS PARA PREVISÃO DE CRISES FINANCEIRAS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias da Universidade de Caxias do Sul.

Orientador: Prof. Dra. Carine Webber

CAXIAS DO SUL

2021

MATHEUS SAUTHIER

MODELOS PREDITIVOS PARA PREVISÃO DE CRISES FINANCEIRAS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação na Área do Conhecimento de Ciências Exatas e Engenharias da Universidade de Caxias do Sul.

Aprovado em 30/06/2021

BANCA EXAMINADORA

Prof. Dra. Carine Webber
Universidade de Caxias do Sul - UCS

Prof. Dr. André Gustavo Adami
Universidade de Caxias do Sul - UCS

Prof. Dr. André Luis Martinotto
Universidade de Caxias do Sul - UCS

Este trabalho é dedicado aos meus pais, por tornarem isto possível.

AGRADECIMENTOS

Gostaria de agradecer em especial aos meus pais e avós, que sempre foram a base e espelho para que eu pudesse persistir e chegar até aqui.

À minha orientadora profª. Dra. Carine Geltrudes Webber, por ter me dado todo o suporte, por me incentivar, animar, direcionar e sempre estar disposição com o esclarecimento de dúvidas e questionamentos.

A todos que direta ou indiretamente contribuíram para que este trabalho fosse possível, em especial aos economistas austríacos que me inspiram a estudar as áreas da economia atreladas às engenharias. Meu mais sincero obrigado!

"Ideias e somente ideias podem iluminar a escuridão."

Ludwig von Mises

RESUMO

A Computação é uma das áreas do conhecimento que se ocupa do desenvolvimento de métodos e técnicas para a construção de sistemas preditivos. Um sistema preditivo parte do reconhecimento de padrões existentes nos dados para prever e antecipar tendências futuras. Neste contexto, o objetivo deste trabalho consiste em analisar dados oriundos da área financeira, tais como movimentações de ciclos econômicos, utilizando técnicas de aprendizagem de máquina para fins de predição. A construção de modelos preditivos na área financeira é um desafio para governos e instituições, uma vez que ela envolve fatores humanos e eventos internacionais. Por outro lado, crises financeiras são acontecimentos raros, logo, difíceis de serem preditos. Contudo, possuem um impacto considerável, que merece ser estudado a fim de ser predito e remediado. Este trabalho partiu de uma revisão sistemática, a partir da qual alguns projetos relacionados foram estudados. Os trabalhos relacionados identificados propõem modelos para predição a partir dados financeiros para problemas de gestão de portfólio, risco de falências, quebra da bolsa de valores e ciclos econômicos. Os modelos relatados são estudos e pesquisas, necessitando ainda de aprimoramento. Na ausência de uma abordagem definitiva para o problema da predição de crises financeiras, propõe-se neste trabalho de conclusão investigar métodos de aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*) aplicáveis a ele. Em termos de conjuntos de dados, existem poucos disponíveis para este fim. O conjunto de dados que será utilizado neste trabalho é a base de dados denominada Macrohistória Jordà-Schularick-Taylor. Como ferramentas de software, selecionou-se a linguagem *Python* e suas bibliotecas, amplamente utilizadas em Ciência de Dados.

Palavras-chaves: Sistemas Preditivos. Crises Financeiras. Economia. Aprendizado de Máquina.

ABSTRACT

Computer Science is one of the areas of knowledge that deals with the development of methods and techniques for building predictive systems.A predictive system consists of recognizing the existing patterns in the data to predict and anticipate future trends. In this context, the aim of this work is to analyze data from the financial area, such as business cycle movements, using machine learning techniques for forecasting purposes. Building predictive models in the financial area is a challenge for governments and institutions, since it involves human factors and international events. On the other hand, financial crises are rare events, and therefore difficult to predict. However, they have a considerable impact, which deserves to be studied in order to be predicted and remedied. This work was based on a systematic review, from which some related projects were studied. The models proposed by the related works for predicting from financial data for portfolio management problems, bankruptcy risk, stock exchange collapse and business cycles. The models reported are studies and researches, which still need improvement. In the absence of a definitive approach to the problem of forecasting financial crises, it is proposed in this work to investigate machine learning and in-depth learning methods applicable to it. In terms of data sets, there are few available for this purpose. The dataset that will be used in this work is the database called the Jordà-Schularick-Taylor Macrohistory. As software tools, Python and its libraries, widely used in Data Science, were selected.

Keywords: Predictive Systems. Financial Crisis. Economy. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As duas grandes emoções que movem os mercados, finanças e investimentos e, portanto, todas as esferas da economia envolvidas, são o medo e a ganância (COMINCIOLI *et al.*, 1995). Nesse contexto, os mercados de ações podem ser utilizados como um dos indicadores dos rumos da economia. No entanto, para sustentar essas afirmações é importante distinguir ciclos econômicos de flutuações econômicas comuns. As pessoas buscam prever continuamente as mudanças para adiantar-se a elas da melhor maneira que podem, sendo por experiências passadas, conselhos ou com auxílio de maneiras mais arrojadas utilizando tecnologias como métodos preditivos. Do ponto de vista de um empreendedor, aqueles que obtiverem as previsões mais assertivas (e executá-las) terão mais sucesso na lucratividade; enquanto os que fracassam na previsão dos seus julgamentos ficam pelo caminho, como punição. O resultado é que os empreendedores de sucesso na livre competição serão aqueles com maior capacidade e melhores ferramentas que possibilitarem antecipar indicadores às futuras condições econômicas.

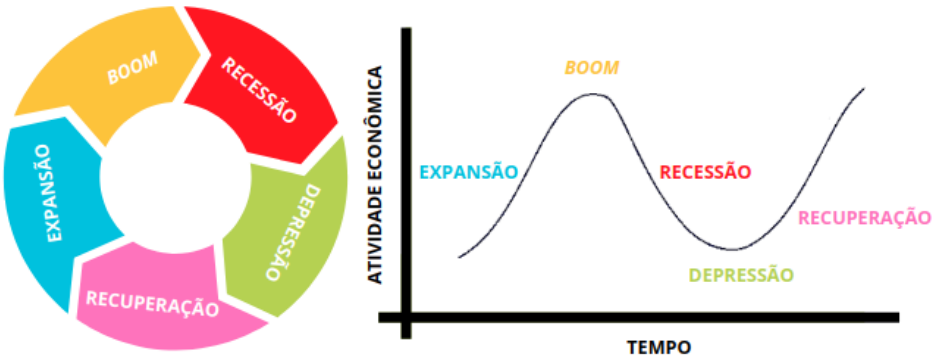
No entanto, uma previsão em um universo gigantesco com inúmeras variáveis, nunca será perfeita, e os empreendedores continuarão a acertar e errar nesse meio, sendo um desafio constante de aprendizado. Assim, há mudanças, continuamente, em todos os domínios da economia: preferências dos consumidores, proporções de investimentos e consumo; forças de trabalho se alteram em termos de quantidade, qualidade, região e tecnologia; técnicas e recursos naturais podem ser descobertos, aprimorados ou chegar ao fim; novas demandas podem surgir, alterando possibilidades de produção; alterações no clima, fauna e flora podem influenciar colheitas etc. Se não fosse assim, jamais haveria qualquer lucro ou perda nos negócios, e não conseguiríamos conceber uma sociedade com mudanças, em que nenhum dado econômico se alterasse (MISES, 2017). Podemos, portanto, esperar flutuações econômicas o tempo todo, pois esta permeia toda nossa vida cotidiana. Estendendo esse raciocínio, a partir do autor (ROTHBARD, 2012), é possível concluir que é um equívoco atribuirmos a depressão econômica a uma esfera específica da economia, a exemplo de "uma queda na demanda por imóveis", ou "uma depressão no setor agrícola".

Segundo economistas austríacos (MISES, 2017; ROTHBARD, 2012; HAYEK, 2011), o problema dos ciclos econômicos advém, na realidade, do *boom* generalizado (ganância) seguido de uma severa depressão (medo). Logo, de acordo com esses autores, uma das questões que uma teoria econômica de uma crise precisa explicar é: por que, subitamente, a atividade econômica que seguia tranquila, com a maior parte dos empreendimentos e investimentos obtendo lucros, de repente, sem qualquer sinalização aparente, sofre mudanças nas condições e apresentam grandes perdas (factíveis a falências) na maioria dos negócios? (ROTHBARD, 2012). Inevitavelmente, revelam-se graves erros nas previsões (variáveis não contabilizadas, mudanças no comportamento generalizado, balanços mal interpretados, expansão de crédito exagerada, au-

mento de impostos, leis protecionistas etc). Considerando-se a atividade empreendedora que está frequentemente realizando previsões, pode-se direcionar a uma possível justificativa: os empreendedores estão constantemente investindo e pagando custos no presente, na expectativa de que vão obter um resultado mais satisfatório no futuro. Porém, em um período de investimentos equivocados, o *boom*, é onde os empresários se iludiram na expectativa de realizar ganhos a partir do crédito bancário facilitado (MISES, 2017) ou decisões governamentais (HAYEK, 2011).

Assim, tais acontecimentos, que impactam várias camadas da economia simultaneamente - são o princípio do ciclo econômico, o qual constitui-se pelas seguintes fases, conforme a Figura 1: a expansão acelerada sem justificativa, marcada por grandes acontecimentos e por mal-investimentos; o *boom* onde todos estão animados e gananciosos; a crise, que chega quando a expansão se reduz e as más decisões tornam-se evidentes; a depressão caracterizada por falências, desemprego e enfim o medo; e a recuperação depressiva, processo necessário de ajuste por meio ao qual a economia retoma as maneiras mais eficientes de satisfazer os desejos de quem quer consumir (ROTHBARD, 2012).

Figura 1 – Ciclo Econômico



Fonte: O Autor (2021)

Como consequência de uma crise financeira, pode-se observar sinais como o desemprego e a falência de empresas (ROTHBARD, 2012). Tais sinais são inevitáveis em toda depressão. Mas há também uma sinalização de esperança em meio a este período de medo: essas ocorrências não são nada mais do que ajustes temporários, uma vez que novas boas decisões forem tomadas, a retomada dos níveis anteriores da economia será rápida.

Nesse contexto, técnicas, ferramentas e *softwares* voltados a análise de dados podem auxiliar na previsão de crises. Na realidade, os autores OZBAYOGLU; GUDELEK; SEZER (2020) destacam que essas análises de dados financeiros para problemas preditivos são um dos temas mais estudados entre pesquisadores. Para evitar riscos desnecessários, economistas, cientistas de dados, engenheiros e empreendedores, fazem uso de métodos preditivos partindo de técnicas de *machine* e *deep learning*. Assim, os modelos preditivos podem possuir maior taxa de acerto

quando aplicadas aos conjuntos de dados atrelados às suas bases de conhecimento e habilidades. Com treinamento extensivo de métodos preditivos, pode-se entregar aos usuários dessas técnicas respostas a perguntas/problemas com maior rapidez, facilidade e assertividade. Computacionalmente, as máquinas que rodam os algoritmos com métodos preditivos irão retornar com análises a *priori* e a *posteriori* com base em grandes repositórios de informações (*dataset*). Os algoritmos voltados para predição (ou seja, que possuem métodos preditivos) podem ser capazes de auxiliar nas decisões e remediar possíveis problemas na compreensão de cenários econômicos, gestão de portfólio, prevenção de perdas, fraudes entre outras aplicações para que possam diminuir riscos dessas atividades.

Ao organizar padrões de dados financeiros de acordo com regras pré-determinadas, as ferramentas que implementam métodos preditivos podem ser capazes de auxiliar em tomadas de decisões, preferências e tendências. Como exemplo, pode-se citar os autores KRAUSS; DO; HUCK (2017), como referências nestas áreas. O referido trabalho apresenta técnicas de *Machine* e *Deep Learning* para a previsão da performance das ações que constituem o índice americano de ações S&P 500 no período de 1992 a 2015. Por se tratar de uma tarefa desafiadora, uma vez que a economia é um campo da sociedade em que sofre constantes mudanças que impactam no resultado de uma previsão, a continuidade de estudos desta natureza é relevante. Para completar, os impactos econômicos são sentidos e refletidos na vida, na sociedade, nas formas de produção. Portanto, a busca por modelos computacionais é um dos desafios da área.

1.1 OBJETIVOS

O objetivo deste trabalho é avaliar métodos computacionais para construção de modelos preditivos de crises financeiras.

Os objetivos específicos são os seguintes:

1. Identificar os métodos computacionais preditivos mais utilizados para predição de crises financeiras.
2. Aplicar os métodos mais utilizados e promissores para um conjunto de dados sobre crises financeiras.
3. Apresentar modelos preditivos para crises financeiras, considerando um conjunto de dados.
4. Avaliar e comparar os resultados obtidos com os trabalhos da área.

1.2 ESTRUTURA DO TRABALHO

Este trabalho está estruturado, definindo modelos preditivos partindo de sua conceituação, diferenciando os modelos de *machine learning* (métodos supervisionados, não supervi-

sionados, semi-supervisionado e por reforço), *deep learning* (sub-área de *machine learning*) e apresentando suas soluções para cada grupo característico. Em seguida, apresenta-se os modelos preditivos, bem como aplicações, para área financeira a partir das contribuições acadêmicas investigadas na revisão sistemática, bem como as métricas de avaliação para definir o desempenho dos modelos. Na seção subsequente, apresentam-se detalhes da metodologia proposta para aplicação do estudo, conjunto de dados, plataformas e bibliotecas que foram utilizadas para aplicação prática do trabalho. Após, apresentamos os resultados de cada passo realizado pela metodologia seguida neste trabalho, desde o pré processamento dos dados selecionados até a obtenção dos resultados a partir das métricas de avaliação. Finalmente apresenta-se a conclusão do trabalho de pesquisa realizado.

2 MODELOS PREDITIVOS PARA A ÁREA FINANCEIRA

Neste capítulo apresenta-se a fundamentação teórica que descreve o funcionamento e quais são os mecanismos mais utilizados na área financeira. Inicialmente, parte-se do conceito de métodos preditivos. Em seguida, conceitua-se *machine learning* e suas várias técnicas de aplicação. Nas seções seguintes são apresentados o conceito de *deep learning* e algumas soluções que constam da literatura na área. Após, métodos preditivos para a área financeira são demonstrados, seguidos pelos trabalhos relacionados. Finalmente, são realizadas as considerações finais do capítulo.

2.1 MODELOS PREDITIVOS

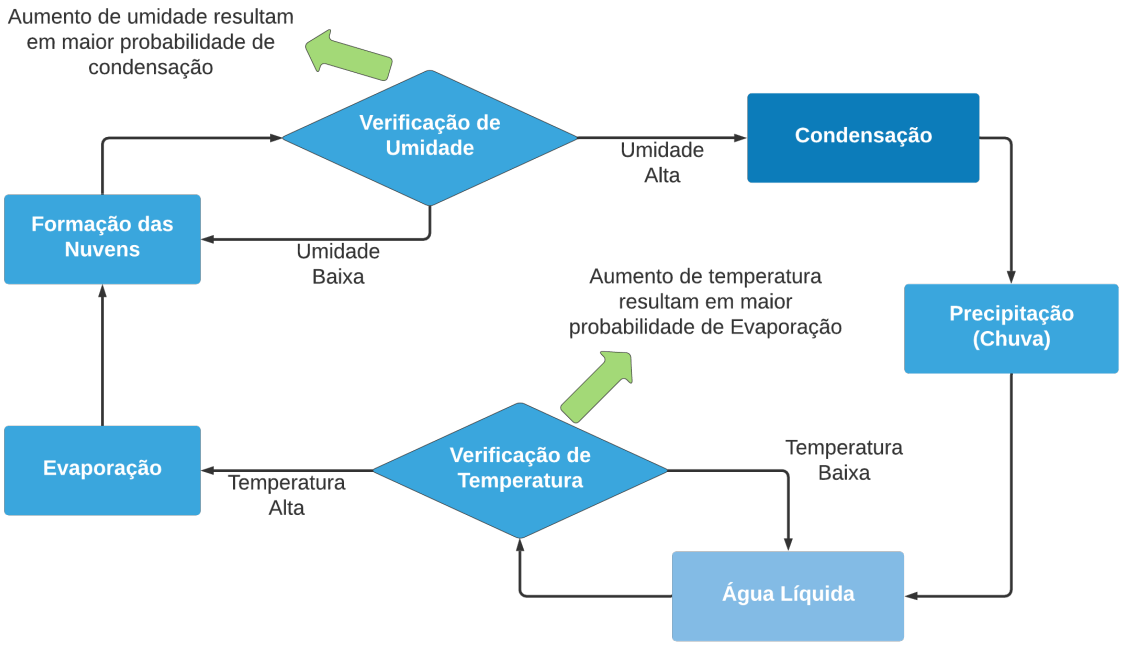
Um sistema preditivo consiste em um agrupamento de regras que por meio de funções matemáticas em uma série de etapas é capaz de realizar uma previsão, identificando padrões quando aplicados a um conjunto de dados. Ao organizar esses padrões de acordo com regras pré-determinadas, o modelo serve para auxiliar em tomadas de decisões, mensurar atividades humanas, preferências e tendências. Consequentemente, essas decisões podem se tornam mais eficientes ao longo do tempo por serem realizadas de acordo com um cenário de necessidades específicas e serem alimentadas cada vez com mais dados e aprendizados passados. Em suma, uma análise preditiva utiliza dados para identificar padrões e, assim, realizar previsões, ou ainda, indicar comportamentos e saídas mais adequadas para uma situação.

Computacionalmente pode-se construir modelos preditivos por meio das técnicas de *machine learning*. Para isso, parte-se de um conjunto de dados, que constituem uma amostra descritiva (exemplos, casos, projetos, períodos, imagens, etc). Em seguida, aplica-se métodos ou algoritmos baseados em estatística, probabilidade, modelos biológicos, sociais, entre outras regras, que buscam localizar, identificar e representar padrões que se repetem nos dados. Tais padrões são o ponto de partida para a construção de um modelo preditivo (RUSSELL; NORVIG, 2002). Como exemplo de um padrão pode-se considerar a associação entre umidade e temperatura para previsão da chuva, conforme o ciclo ilustrado na Figura 2. Os padrões são descobertos por meio de associações dos valores de atributos (umidade alta ou baixa, temperatura alta ou baixa, etc) de um conjunto de dados (*dataset*) em relação a um atributo alvo ou classe (evaporação ou água líquida; formação de nuvens ou condensação).

2.2 *MACHINE LEARNING*

Machine Learning ou em português aprendizado de máquina, é uma técnica de inteligência artificial onde é possível aplicar o conceito de modelos preditivos. Com os modelos preditivos, pode-se introduzir as técnicas de aprendizagem de máquina, cada uma com a ta-

Figura 2 – Associação de umidade e temperatura para previsão de chuvas



O Autor (2021)

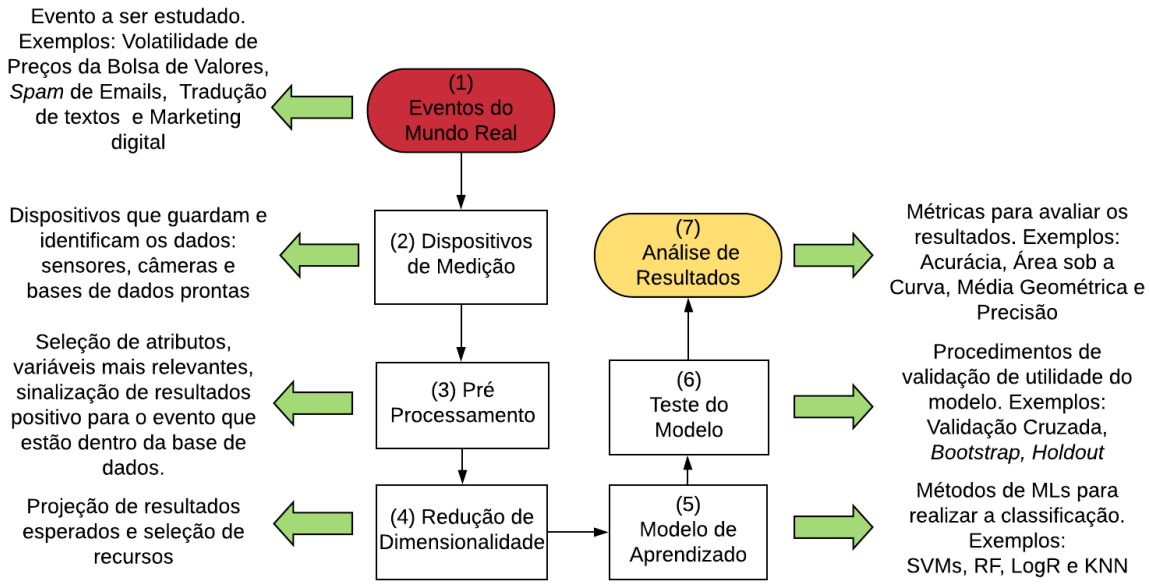
refa simples de aprendizagem para realizar uma determina predição (RUSSELL; NORVIG, 2002). Esses modelos são estruturados em computadores onde, a partir de treinamentos, são capazes de retornar respostas esperadas por meio de associações de dados. Os computadores aprendem através de um repositório de informações (chamado de *dataset*), os quais podem ser imagens, números, associações de palavras, instruções e enfim qualquer observação de dados que a tecnologia possa identificar para resultar em um modelo de aprendizado. Suas aplicações abrangem uma ampla área como previsão do tempo, detecção de *spam*, visão computacional, análise de sentimento, reconhecimento de padrões e entre outras soluções para problemas de predição.

Uma vez que o algoritmo de aprendizado de máquina aprendeu, o mesmo poderá ser capaz de executar tarefas complexas e dinâmicas, possuindo maior precisão de predição, ter uma capacidade de reação assertiva e de maneira mais inteligente, podendo modificar seu comportamento com intervenção humana mínima conforme novas experiências são utilizadas para treinar o algoritmo (TAULLI, 2020). Essa alteração de comportamento ocorre fundamentalmente devido ao estabelecimento de regras lógicas (parâmetros do modelo preditivo) que visam melhorar o desempenho da aplicação.

Para compreender a complexidade de um sistema de aprendizado de máquina, fragmenta-se todos os passos de aplicação de um modelo preditivo, onde em cada fase temos (1) eventos do mundo real; (2) dispositivos de medição e captação dos dados; (3) pré processamento de dados; (4) redução de dimensionalidade; (5) aplicação do modelo preditivo; (6) teste do modelo preditivo; e finalmente (7) resultados da classificação realizada pelo modelo. Na Figura 3 pode-

se visualizar o fluxograma que ilustra o processo de solução do método por aprendizagem de máquina. Ademais, ressalta-se que existem quatro abordagens de *machine learning*: aprendizagem supervisionada, aprendizagem não supervisionada, aprendizagem semi-supervisionada e aprendizagem por reforço, que serão discutidos nos próximos parágrafos desta seção.

Figura 3 – Fluxograma de funcionamento de *Machine Learning*



O Autor (2021)

Nos métodos preditivos que fazem parte do grupo de aprendizagem supervisionada, os rótulos dos dados estão estabelecidos, ou seja, existem resultados esperados para os dados que serão compilados na máquina. Fundamentalmente, a aprendizagem pode ser estimada a partir de atributos reais ou valores dentro de um conjunto finito. Para os autores RUSSELL; NORVIG (2002), o modelo preditivo que utiliza aprendizagem supervisionada procura relacionar pares entre as entradas (dados do problema) e saídas (resultados esperados), buscando aprender uma função que faz o mapeamento desta entrada e saída. Essa classificação é destinada para encontrar padrões que possam ser organizados em um processo analítico, a exemplo de classificar 3 espécies de flores a partir dos atributos dispostos em um *dataset*, como cor, forma e textura, (SHUKLA *et al.*, 2020). No exemplo, o modelo preditivo dessa abordagem trabalha com algoritmos capazes de analisar as características de cada flor registrada na base de dados e predizer a chance de uma determinada flor, ainda não rotulada, de participar de uma determinada espécie (rótulo). Desse modo, o computador é capaz de tomar decisões precisas quando recebe novos dados não rotulados a partir de um treinamento com dados que possuem rótulos conhecidos.

Ainda, em outro estudo, os autores destacaram a vantagem do uso de aprendizagem supervisionada para análises forenses ao vivo em plataformas IoT, uma vez que é possível forne-

cer uma previsão com base em ocorrências anteriores (KEBANDE *et al.*, 2020). Nesta abordagem, destaca-se um cenário hipotético (incidente em um porto que produziu um grande estrondo). Um usuário solicita o gerenciamento de incidentes forenses dinamicamente automatizado para analisar o "incidente em potencial". Forma-se assim, um relatório a partir dos equipamentos localizados em torno a região sinalizada, como sensores e câmeras que enviam os sinais captados para o servidor de aplicação. No servidor, técnicas de aprendizagem supervisionadas são utilizadas nos dados do relatório. Assim, os dados do relatório são comparados a base já rotulada, permitindo classificar o incidente (por exemplo, tiroteio, acidente etc.). O agente solicitante então pode utilizar os resultados para tirar conclusões que ajudam na formação de uma hipótese sob um incidente potencial. Destacam-se como alguns métodos do aprendizado supervisionado os algoritmos de máquinas de suporte vetorial, florestas aleatórias e regressão logística.

Na segunda abordagem, a aprendizagem não supervisionada, a base de dados utilizada não possui rótulos, ou seja, não há padrões definidos para os agrupamentos da variável resposta (RUSSELL; NORVIG, 2002). Os modelos que fazem parte deste tipo de aprendizagem, conduzem de modo interativo a análise da base de dados sem intervenção humana, ou seja, não fazem uso da informação dada por cada rótulo ao qual cada dado pertence. Em resumo, o modelo preditivo aplicado aprende os padrões de entrada, embora não é fornecido nenhum *feedback* explícito se tais padrões aprendidos estão corretos. Os autores RUSSELL; NORVIG (2002), identificam que a tarefa mais comum deste tipo de aprendizagem é o agrupamento: detectar grupos exemplos (rótulos) na entrada que serão potencialmente úteis para a saída. Para LIEBER *et al.* (2013), tais modelos preditivos são frequentemente usados como uma primeira etapa na mineração de dados para se obter as primeiras impressões de padrões relacionados em conjuntos de dados complexos. Denomina-se processo de *clustering* o método que consiste em agrupar dados, fornecendo uma visão sobre as suas similaridades, por meio da formação de grupos (*clusters*). Como algoritmos de aprendizagem não supervisionada pode-se citar os particionais, hierárquicos e por densidade. Para os algoritmos particionais divide-se o *dataset* em grupos onde a quantidade de grupos é determinada pelo usuário. Esse algoritmo seleciona objetos (divididos de acordo com a medida de similaridade adotada) como centros dos conjuntos, de modo que cada objeto que fique no *cluster* forneça o menor valor distância entre o objeto e o centro do mesmo. Como exemplo, os modelos preditivos conhecidos como *k-means* caracterizam esse grupo, e é um dos que se obteve maior eficiência em um grande número de aplicações segundo a literatura (OLIVEIRA, 2018).

Os algoritmos hierárquicos tem sua organização de dados em uma estrutura hierárquica de acordo com a proximidade entre os objetos, onde seus resultados são geralmente demonstrados como uma árvore que divide a sua base de dados em conjuntos menores. O resultado pode ser obtido a partir de recortes em diferentes níveis de acordo com o número de conjuntos desejados. Esta forma de representação de resultados fornece descrições informativas para as estruturas de grupos (*clusters*) em potencial, principalmente quando há relações hierárquicas nos dados como, por exemplo, dados de pesquisa sobre evolução de espécies. Todavia, os al-

goritmos hierárquicos (e também os particionais) podem encontrar dificuldades para descobrir *clusters* de formas arbitrárias (CARLANTONIO, 2001). Como alternativa para esta dificuldade, têm-se os algoritmos baseados em agrupamentos por densidade. Esses algoritmos definem os *clusters* como regiões densas, separadas por regiões menos densas que representam os ruídos no espaço amostral. Essas concentrações de elementos podem estar distribuídas arbitrariamente em cada agrupamento e por isso, os algoritmos baseados em densidade são adequados para descobrir *clusters* com formatos aleatórios. Esses algoritmos diferem-se pelo processo de formação: uns determinam os *clusters* de acordo com a densidade da vizinhança dos objetos, outros, trabalham de acordo com alguma função matemática de densidade.

Em terceiro lugar, a aprendizagem de máquina semi-supervisionada, pode ser definida a partir situações em que a base de dados se divide em duas parcelas: (1) dados rotulados, onde os rótulos estão classificados com as regras de normalidade[1] e anomalias[2]; e (2) dados não rotulados, que por sua vez possuem exemplos de casos recorrentes, podendo inferir a separação entre as classes de amostras (OLIVEIRA, 2018).

Pode-se identificar métodos preditivos de aprendizagem semi-supervisionada onde o problema possui dados onde poucos são rotulados e deve-se fazer o que puder de uma grande coleção de exemplos não rotulados (RUSSELL; NORVIG, 2002). Esses autores também destacam que mesmo os rótulos em si podem não ter os resultados que esperamos, como por exemplo, tentar classificar idades a partir de fotos de pessoas: pode-se reunir alguns exemplos rotulados tirando fotos das pessoas e perguntando a idade (aprendizagem supervisionada). Mas, algumas das pessoas mentiram sua idade, fazendo com que não exista apenas ruído aleatório nos dados, mas as imprecisões nos próprios rótulos. Descobrir as imprecisões é um problema de aprendizagem não supervisionada, envolvendo imagens, idades relatadas e idades (desconhecidas) verdadeiras. A partir daí, tanto ruído como falta de rótulos cria um espaço entre aprendizagem supervisionada e não supervisionada. Para aplicação, primeiramente é ensinado ao computador (utilizando aprendizagem não supervisionada) como os atributos de normalidade se apresentam, a partir dos dados não rotulados. Assim, o método preditivo passa a associar cada observação com uma pontuação à probabilidade do dado participar da região de normalidade. Em seguida, são utilizados os dados rotulados (com exemplos de normalidade e anomalias) para definir um limar na pontuação obtida, ou seja, produzir uma avaliação no método preditivo de detecção. A vantagem do uso aprendizagem semi-supervisionada é que essas possuem excelente desempenho em cenários onde há poucas amostras anômalas, como observado por FACURE (2017) que utilizou esse subtipo de *machine learning* para detecção de fraudes. Neste estudo, a base dados de transações feitas por cartão de crédito por consumidores europeus, foram repartidas em 3 subamostras onde 80% foi destinado para treino dos modelos e as outras duas foram separadas igualmente em observações normais e anormais, onde uma foi utilizada para ajustar o limiar da

[1] Regras de normalidade: Dados que estão dentro do padrão esperado no universo atribuído.

[2] Regras de anomalias: Dados que estão fora da combinação da noção bem definida de normalidade, apresentam resultados distintos das observações normais.

pontuação da anomalia e outra para testar a performance das técnicas consideradas. As soluções de mistura de gaussianas e modelo gaussiano apresentaram os melhores desempenhos nas métricas de performance. Em suma, o aprendizado semi-supervisionado é quando utilizam-se dados com e sem rótulo em qualquer instante do projeto de um modelo, (LELIS, 2007).

Finalmente, a aprendizagem por reforço é conhecida como um modelo de aprendizado comportamental. O agente (modelo preditivo) aprende a partir de uma série de reforços: recompensas e punições. Em outras palavras agente recebe um "*feedback*"do processamento de dados, direcionando para o sistema a melhor decisão. O exemplo que os autores RUSSELL; NORVIG (2002) nos trazem é a de uma gorjeta ao final de uma corrida: o agente pode indicar o motorista de que algo saiu errado (não dando gorjeta) ou que a viagem atendeu as expectativas (dando gorjeta). Enfim, cabe ao agente decidir qual das ações anteriores ao reforço foram as maiores responsáveis por isso. Em outro exemplo, pode-se observar as ações em um jogo de computador, onde o agente (personagem) está inserido em um ambiente (mapas) que são descritos por um estado (outros personagens, localização das missões, atributos do personagem etc) que representam o momento do ambiente em que o personagem está inserido. Durante o decorrer do jogo são realizadas ações (andar para direita ou para esquerda; finalizar uma das tarefas; comprar novos equipamentos; e descobrir novos mapas) onde cada ação existe uma recompensa (avanço no jogo) ou punição (personagem retornar ao ponto inicial) que são reforçadas conforme cada decisão tomada. O personagem sempre estará inserido em um estado, decidindo por uma ação dentre as possíveis mudando de estado, resultando em um tipo de reforço (recompensa ou punição). Assim, na aprendizagem por reforço, os métodos preditivos buscam ao longo do tempo, tomar as melhores decisões baseados nas recompensas e punições recebidas no *feedback* passado.

Este subtipo de *machine learning* se difere dos outros uma vez que suas soluções não são treinadas com o conjunto de dados de amostragem, mas em um ambiente dinâmico através de interações do tipo "tentativa e erro". Ao contrário do outros métodos de aprendizado, não existem pares de entradas e saídas (ou agrupamentos) para serem utilizados no treinamento. Para tomar uma ação, o modelo preditivo imediatamente recebe uma valoração (recompensa) da sua decisão mas não obtém uma resposta direta de qual deveria ser a melhor ação para atingir o objetivo. Uma sequência de decisões bem-sucedidas por parte do modelo irão resultar no processo sendo reforçado, pois ele tende a entregar o melhores desempenhos para os problemas ao longo do tempo. A aprendizagem por reforço, portanto, precisa obter experiência dos possíveis estados, classificações, ações e recompensas para atingir o melhor desempenho. Em alguns casos observa-se que os tipos de aprendizado pode ser combinados. O estudo desenvolvido SOLEYMANI; PAQUET (2020), propôs uma estrutura de aprendizado por reforço, utilizando uma rede neural convolucional. Nele, foi discutido o problema de realocação contínua de ativos financeiros, utilizando modelos preditivos de aprendizado de máquina para minimizar o risco no gerenciamento de um portfólio.

Na próxima seção, são mostradas as principais técnicas de *Machine Learning* de acordo com seu subtipo de aprendizado. Essas técnicas são brevemente exemplificadas, onde poderão ser utilizadas para a continuidade do presente trabalho.

2.3 PRINCIPAIS TÉCNICAS DE *MACHINE LEARNING*

Existem diversas técnicas algorítmicas (modelos preditivos) que se aplicam ao problema de aprendizagem de máquina. Propõe-se, neste estudo, uma breve descrição dessas técnicas mais comuns conforme o processo de revisão sistemática descrito nas próximas seções. Descrições complementares sobre estas técnicas podem ser encontradas em diversas obras da literatura da área (LUGER, 2013; COPPIN, 2010; MITCHELL, 1997).

2.3.1 Máquinas de Suporte Vetorial

Uma máquina de suporte vetorial (SVM) é uma técnica de aprendizagem supervisionada que tem por objetivo desenvolver uma classificação, para novas entradas de dados. Tais entradas são designadas aos possíveis grupos em que cada dado faz parte, usando como base a análise de regressão[3]. Um grupo pode pertencer a várias categorias, onde a máquina constrói um modelo que atribui os novos dados a cada categoria. O modelo preditivo SVM é desenvolvido a partir das entradas de treinamento, mapeando esses dados no espaço e utilizando a análise de regressão para encontrar um hiperplano[4]. Após ser treinando, o modelo será capaz de redirecionar as novas entradas em relação ao hiperplano e classificar entre as suas categorias. Para RUSSELL; NORVIG (2002) destaca-se que o SVM possui três propriedades essenciais: (1) constrói um separador (um limite de decisão entre as classes); (2) criam uma separação linear em hiperplano, separando em várias classes conforme o problema exigir; e por fim (3) mantêm os exemplos de treinamento podendo precisar armazenar todos eles. Por exemplo, deseja-se aplicar o SVM a duas entradas, que irá retornar uma única saída onde são classificados os pontos em um gráfico pertencentes a duas categorias. Neste contexto, pode-se visualizar a classificação do *dataset* na Figura 4. Os círculos azuis indicam os pontos de dados que pertencem a categoria 1, e as cruzes vermelhas pertencem a categoria 2. Cada dado representado no gráfico (círculo ou cruz) possui um ponto cartesiano mapeado no espaço bidimensional (x,y). O SVM utiliza algoritmos de regressão para encontrar um hiperplano (linha preta). Por sua vez, esta linha tem por objetivo separar os dados entre as duas categorias, onde essa linha é utilizada para classificar as novas entradas entre círculos azuis e cruzes vermelhas.

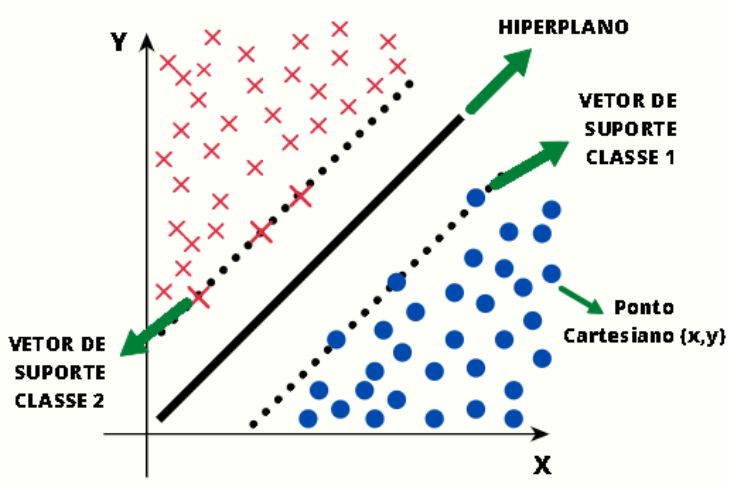
As máquinas de suporte vetorial tem a vantagem que reduzir a possibilidade de *overfitting*[5], CHATZIS *et al.* (2018). Conforme o estudo dos autores, salienta-se a desvantagem do

[3] Regressão: relação entre as variáveis. As equações de regressão podem ajudar a prever os resultados com base nas estradas dos computadores de treinamento.

[4] Hiperplano: Superfície no espaço de k dimensões onde são separadas as metades do espaço.

[5] *Overfitting*: ocorre quando o modelo testado no dados de treino tem um desempenho muito bom mas quando

Figura 4 – Exemplo de classificação por máquinas de suporte vetorial



O Autor (2021)

SVM no fato de que ele pode limitar o seu potencial em análises que exigem profundidade na classificação, uma vez que se constituem em uma caixa-preta, ou seja, seus detalhes de implementação estão "escondidos", fazendo com que alguns resultados dificultem a interpretação. Em outra abordagem, alguns autores compararam o desempenho das máquinas de vetores de suporte entre outros modelos preditivos, e concluíram que para melhorar o desempenho do método preditivo destacou-se a necessidade de selecionar os melhores atributos para o modelo (PETROPOULOS *et al.*, 2020).

2.3.2 Regressão Logística

A regressão logística (LogR) pode ser observada em cenários que não são os números diretamente quantificados que se fazem necessários para obter bons resultados, mas a probabilidade de ocorrência de um evento (GONZALEZ, 2018). Uma vez que um modelo preditivo linear sempre busca uma previsão completamente confiante de 1 ou 0 (sem valores no intervalo), o LogR busca realizar previsões mais graduais. Conforme RUSSELL; NORVIG (2002), a regressão linear pode resolver esses problemas suavizando a função de limiar[6]. Por exemplo, categorizar variáveis por grupos, onde esses grupos classificam a probabilidade de uma precipitação ocorrer; e assim, as chuvas. A regressão logística usa uma função para prever os valores em um determinado intervalo. Em outras palavras, este modelo preditivo pode ser utilizado em situações em que os atributos são de uma natureza n (sim, não ou talvez; comprar ou não comprar; chover ou não chover). Assim, a regressão logística possibilita estimar a probabilidade de ocor-

treinado com a base teste, seu desempenho é muito inferior em comparação aos dados de treino. Em resumo, o modelo aprendeu tão bem as relações existentes no treino, que acabou apenas “decorando” o que deveria ser feito, ao invés de aprender

[6] Uma função linear através da função de limiar cria um modelo preditivo cujo o classificador é linear: busca-se criar uma linha que melhor se encaixe aos dados de entrada do modelo preditivo, criando uma linha reta entre esses dados.

rência de um evento. A Equação 2.1 exemplifica um modelo de regressão logística que retorna uma classificação entre dois grupos. A variável dependente binária Y, onde pode assumir o valor desses dois grupos (1 ou 0), havendo um conjunto P (ocorrência de um evento) de variáveis independentes. Ainda, $g(x)$ representa a soma de um conjunto de variáveis independentes X multiplicados pelos coeficientes estimados a partir de um conjunto de dados B. Estes coeficientes influenciam no resultado final: para coeficientes positivos (no modelo apresentado), o resultado retorna uma probabilidade maior, enquanto para coeficientes negativos, probabilidade menor de um evento ocorrer.

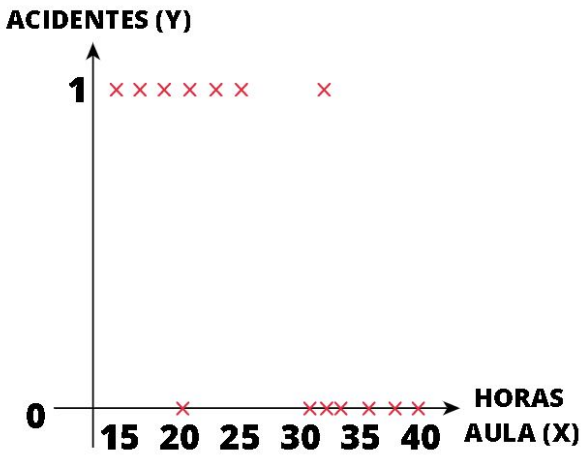
$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}} \tag{2.1}$$

onde,

$$g(x) = B_0X_0 + B_1X_1 + ... + B_nX_n$$

Outro exemplo de aplicação pode ser visto supondo-se que uma agência financeira está decidindo se deve ou não segurar um veículo. Esta agência pode verificar a probabilidade do tomador de crédito se ele bater seu carro a partir do tempo em que esteve na autoescola; onde esses dados são de tempo, mas não se tem conhecimento se este atributo se relaciona com a probabilidade do segurado sofrer um acidente. Partindo-se de um grupo de K pessoas, onde os dados de tempo na autoescola e se elas sofreram acidentes já estão computados, pode-se visualizar que a maioria das pessoas que sofreram acidentes, ficaram pouco tempo na autoescola conforme na Figura 5. O eixo Y está marcado em 1 se pessoa sofreu acidentes e 0 o contrário, enquanto o eixo X representa o tempo que cada condutor passou na autoescola. Assim, cada ponto no plano cartesiano relaciona o tempo de aprendizado na autoescola (X) e se o indivíduo sofreu acidente (Y).

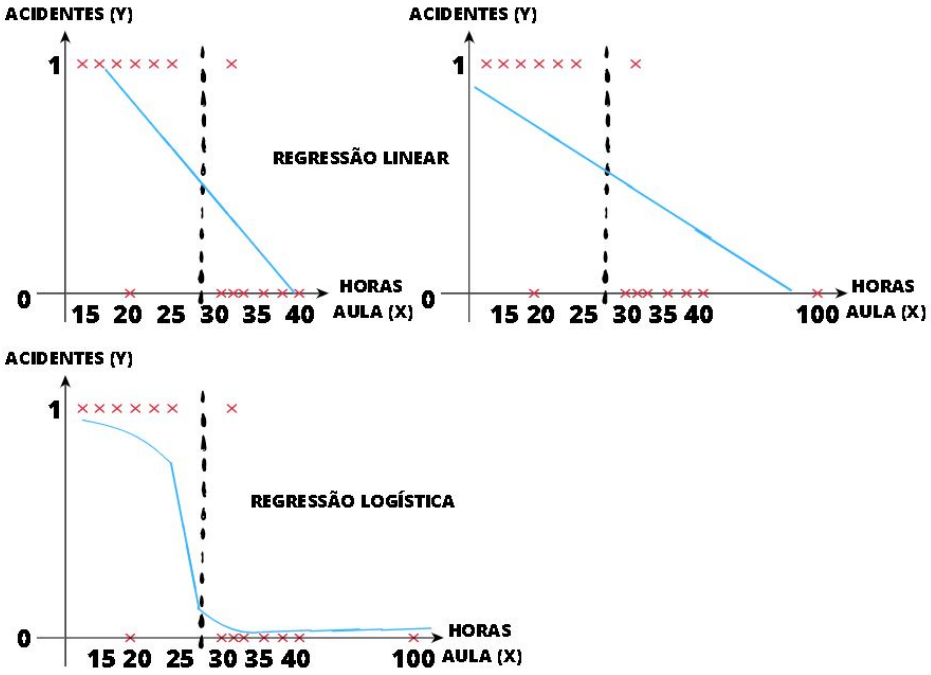
Figura 5 – Relação de acidentes com horas de aula na autoescola



O Autor (2021)

A vantagem dos modelos preditivos de regressão logística é que fornecem uma previsão variante entre 0 e 1, o que permite que tenha a capacidade de uma interpretação probabilística, conforme observado por QU *et al.* (2019). Portanto, ela não será influenciada por novos atributos discrepantes: supondo-se que em um conjunto normal de pessoas que fizeram autoescola entre 10 a 40 horas foi adicionado um único condutor que realizou várias horas de aula a mais (100 horas). Ao contrário do método de regressão linear (que alteraria seu limiar influenciado por novos dados, mudando sua reta), é encontrada uma curva em formato de "S"em que melhor se ajusta o conjunto de dados. Todas essas informações podem ser vistas na Figura 6, a seguir.

Figura 6 – Ajuste Regressão Linear vs Regressão Logística



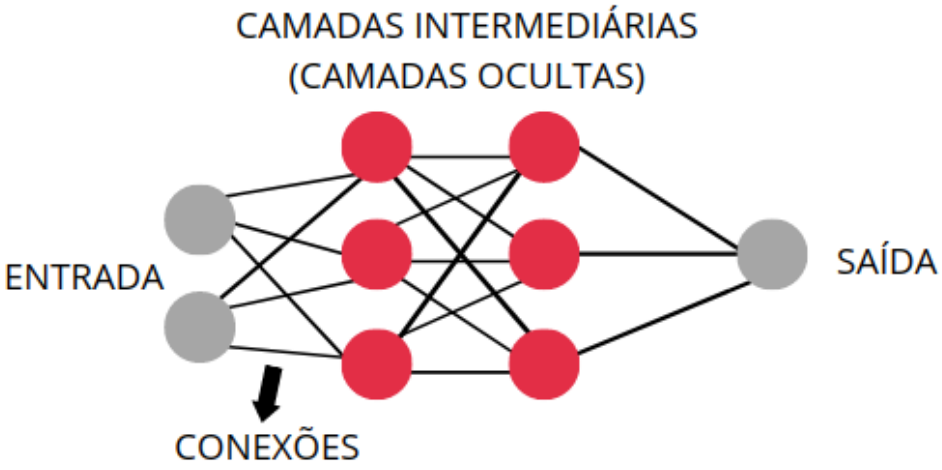
O Autor (2021)

2.3.3 Redes Neurais

As redes neurais (NN) são inspiradas no mecanismo biológico dos neurônios do cérebro humano, onde consistem em várias camadas de neurônios, desde sua camada de entrada até camadas mais ocultas chegando na saída de retorno de resultado. No contexto biológico, entende-se que cada neurônio pode enviar um impulso elétrico (informação) para outro neurônio conectado a partir de suas sinapses. Quando o envio de informação é executado, classifica-se como um neurônio ativado, terminologia utilizada também para as redes neurais artificiais, ou seja, os neurônios estão a todo momento recebendo impulsos elétricos de outros neurônios que por sua vez também estão enviando informações (ou não) para outros. A partir da formação de várias camadas de neurônios conectadas, é possível estabelecer reconhecimento de padrões e obter resultados. A Figura 7 apresenta uma rede neural com todas as suas conexões. Como

exemplo, pode-se utilizar um conjunto de dados onde estão dispostas várias formas escritas de dígitos numéricos. Existem várias formas de escrever tais dígitos; todavia ainda há padrões que possam ser reconhecidos para cada valor do sistema decimal (voltas, número de arestas, formato geométrico, etc). Para encontrar o número 3 por exemplo, a rede neural pode ser utilizada até que se encontre similaridades características que reconheçam o dígito 3 que se busca.

Figura 7 – Rede Neural



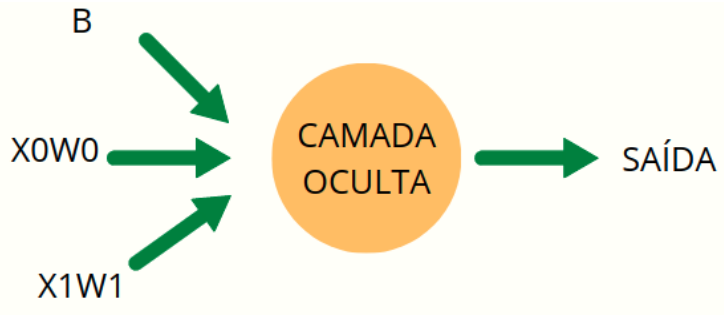
TAULLI (2020)

O modelo NN possui unidades (também chamadas de neurônios ou nós). Uma ligação da unidade i para a unidade j serve para propagar a ativação a de i para j. Conforme RUSSELL; NORVIG (2002), esses nós possuem um peso w, que indica sua relevância do valor na rede que irá para próxima camada. A Equação 2.2 representa o funcionamento matemático das redes neurais: bem como os modelos preditivos de regressão, cada unidade tem uma entrada a com um valor associado a w, onde cada unidade j primeiro calcula a soma ponderada de suas entradas.

$$in_j = \sum_{i=0}^{n} w_{ij} * a_i \tag{2.2}$$

Assim, as redes com alimentação para frente são dispostas em camadas de forma que cada unidade recebe a entrada somente a partir de unidades na camada imediatamente anterior. Redes mais complexas, que possuem várias camadas, têm uma ou mais camadas ocultas que não são conectadas às saídas da rede. Tais camadas ocultas usam funções matemáticas cujo resultado se tornará a saída. Há também uma constante utilizada nos cálculos da função chamada de viés (bias). Em seu nível mais básico, esse tipo de treinamento se movimenta somente uma vez em (1) camada de entrada; (2) camada oculta; e (3) saída, sem iteração de novos ciclos. No entanto, para estruturas de redes neurais com maior número de camadas poderiam cair em novas ramificações de redes neurais, com as saídas se transformando novamente em entradas. A Figura 8 ilustra um nodo de uma rede neural.

Figura 8 – Funcionamento de uma rede neural



TAULLI (2020)

Supondo que o modelo ilustrado na figura 8 é utilizado para prever se o preço das ações de uma empresa vão subir ou cair, é possível verificar o que cada variável de entrada representa, respectivamente: B, viés (bias); $X0$, as receitas da empresa; $W0$, o peso que o atributo $X0$ reflete no modelo; $X1$, a margem de lucro da empresa; $W1$, o peso em que a variável $X1$ representa no cálculo da rede neural. Em seguida, após a definição desses valores (parâmetros), cada neurônio utiliza realizam calculos que irão classificar os valores do *dataset*. A vantagem é que, por conta de geralmente a função não ser linear, os resultados refletem de uma melhor forma o mundo real, uma vez que as informações também não são lineares (TAULLI, 2020). Neste contexto, uma função comum utilizada nesses treinamentos é a sigmoide[7]. Todavia, o modelo apresentado não seria útil em modelos para aplicações complexas. Para adicionar mais complexidade, é necessário que existam várias camadas na rede neural (camadas ocultas), também conhecido com *MultiLayered Perceptron* (Perceptron Multicamadas). Essa abordagem auxilia a usar retro-propagação, que permite que a saída da rede se conecte novamente a entrada para uma nova rodada de cálculos. Outra vantagem das redes neurais é a de que ela é capaz de computar várias funções de verossimilhança[8], podendo resolver problemas de classificação binárias categorizando o conjunto de dados utilizados na entrada. No entanto, existem também desvantagens ao utilizar redes neurais. Uma das principais desvantagens, é o processo de ajustes de pesos no modelo, uma vez que muitas abordagens tradicionais como a inserção de valores aleatórios, demandam muito tempo de processamento (TAULLI, 2020).

Diante deste problema, a retro-propagação se configura como uma solução (TAULLI, 2020). Esta técnica tem o objetivo de ajustar a rede neural quando erros são encontrados e, iterar novos valores na rede novamente (RUSSELL; NORVIG, 2002). O processo de retro-propagação realiza pequenas modificações ao longo da rede, onde a cada iteração o modelo é otimizado. Por exemplo, supondo-se que uma das entradas de dados tem uma saída de 0,6 (em uma escala de 0 a 1). Isso significa que seu erro é de 0,4 e está abaixo da média. Utilizando a ferramenta

[7] Função sigmoide: $f(x) = \frac{1}{1+e^{-x}}$, para todo x real. Ela é a solução da equação $\frac{dy}{dx} = \lambda(y)(1-y)$, com y avaliado entre 0 e 1. A função sigmoide possui amplo uso em campos da economia e computação

[8] Verossimilhança: Compreende-se que varia daquilo que parece intuitivamente verdadeiro, atribuído a uma probabilidade de verdade

de retro-propagação, no entanto, é possível o que a rede neural em uma nova iteração retorne o valor de 0,7 na sua saída e assim sucessivamente até que o valor esteja o mais próximo possível de 1 (TAULLI, 2020).

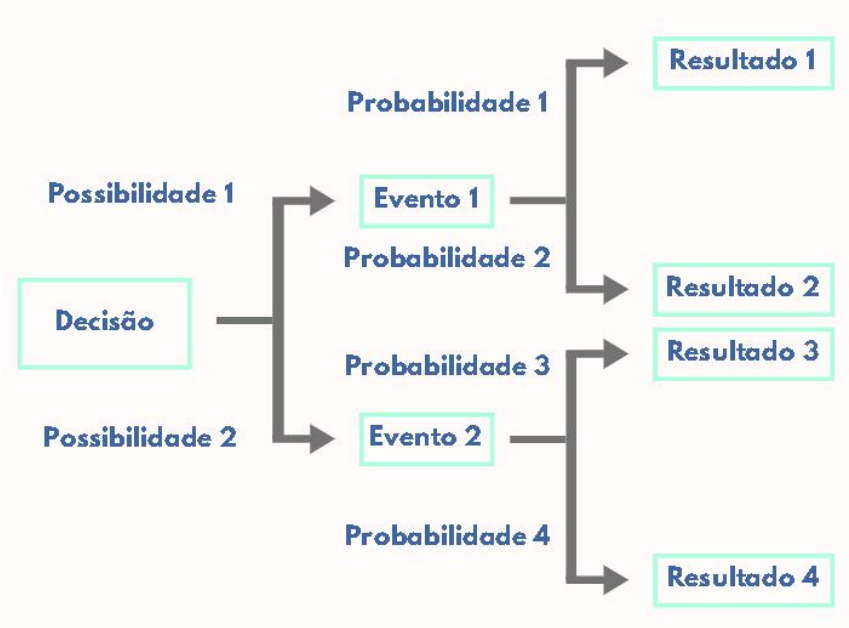
No campo das redes neurais, estudos sugerem que tem havido inúmeras pesquisas sobre aplicações para prever falências financeiras conforme citam os autores QU *et al.* (2019). Nesse contexto, o artigo apresenta sistemas com precisão de 87% de eficiência usando redes neurais com várias camadas ocultas. Outras soluções mais comuns incluem, por exemplo, as redes neurais convolucionais e redes neurais recorrentes em sua variação de memória longa de curto prazo. Esses conceitos são observados nas próximas seções de aprendizado profundo (*deep learning*).

2.3.4 Árvores de Decisão

O modelo preditivo por árvores de decisão (CART) parte inicialmente de um único nó de raiz (também conhecido como nodo) que está no topo de uma estrutura semelhante a de um fluxograma, onde cada estágio dele é outro nodo. Em outras palavras, uma árvore de decisão pode ser representada a partir de uma estrutura com vários nodos que tem uma função que toma como entrada um vetor de valores e retorna uma decisão a partir desses valores, conforme citado por RUSSELL; NORVIG (2002). A partir desse ponto, podem existir diversas ramificações que consistem em caminhos de decisões também chamados de divisões. Uma árvore de decisão alcança sua decisão executando uma sequência de testes. Cada nó interno na árvore corresponde a um teste do valor dos atributos de entrada, e as ramificações dos nós são classificadas com os valores possíveis do atributo selecionado. Ao final da árvore, os nós (conhecidos como nós de folha) especificam o valor a ser retornado. Esse modelo preditivo é comumente utilizado na construção de fluxogramas, uma vez que esses de modo geral permitem criar caminhos de acordo com escolhas (TAULLI, 2020). Ainda, o propósito da árvore de decisão, em computação, é obter diversas divisões capazes criar subconjuntos de dados do *dataset* mais puros. Por sua vez, a pureza dos dados aumenta a medida em que cada nó da árvore contém menos classes (ou apenas uma). Pode-se compreender que uma árvore de decisão possui regras em seus nós, e os nós folhas representam a decisão a ser tomada. Nesses pontos subdivididos, um algoritmo é usado para tomada de decisão e uma probabilidade é calculada. No final da árvore, pode-se observar que os nós folhas estão possuem os resultados, conforme a Figura 9.

Em uma árvore de decisão, uma escolha é tomada através do caminho tomado desde o nó raiz (início) até o nó folha (fim). Para elucidar, pode-se supor a seguinte situação: em um jogo de tabuleiro que depende de um grande arranjo de decisões, chega-se a uma situação em que não se tem claro qual caminho deve-se tomar. Portanto, recorre-se ao processo de árvore de decisão para tomar uma escolha de forma mais assertiva. Partindo do nó raiz, que contém perguntas em torno do jogo, segue-se uma sequência respostas de sim e não de acordo com perguntas de cada nodo até chegar as folhas, onde apresentará o melhor caminho estimado a

Figura 9 – Representação de uma árvore de decisão



O Autor (2021)

partir das respostas realizadas. Assim, a abordagem de árvores de decisão, em geral, possui um melhor desempenho com dados não numéricos. Ainda, são fáceis de compreender e funcionam bem com o grandes conjuntos de dados (TAULLI, 2020).

2.3.5 Florestas Aleatórias

As florestas aleatórias (RF), como o próprio nome sugere, tem por característica criar várias árvores de decisão de maneira aleatória, formando uma floresta de CART, onde cada árvore retorna um resultado. Em outras palavras, as florestas aleatórias são uma combinação de várias árvores de decisão onde cada árvore depende de valores de um vetor de amostras (BREIMAN, 1999). Ao contrário do que ocorre quando é utilizado o modelo de árvore de decisão, as florestas aleatórias, selecionam aleatoriamente os valores do vetor de entrada para cada árvore nos dados de treino, não utilizando toda sua base de treino completa. Inicialmente, parte-se de um processo conhecido como *bootstrap*[9], onde, com essa seleção de amostras, é construída a primeira árvore de decisão. No segundo passo, o modelo preditivo irá escolher aleatoriamente duas ou mais variáveis, as quais então serão realizados cálculos com base na amostragem selecionada para definir qual valor será utilizado no primeiro nó. Desta forma, a árvore será construída até as suas folhas (resultados). Na construção das próximas árvores, os processos anteriores se repetem, mas essas árvores serão diferentes umas das outras, uma vez que na seleção dos valores do vetor o processo acontece de maneira aleatória. Ao longo da criação de novas árvores destaca-se que esta estratégia costuma evitar *overfiting*, pois seus valores de seleção são es-

[9] *Bootstrap*: método de re-amostragem onde as amostras selecionadas podem ser repetidas na seleção

colhidos aleatoriamente de forma iterativa e são utilizados apenas um subconjunto do *dataset* disponível (TECH, 2020). Ainda, conforme FRIEDMAN *et al.* (2001)se faz necessário distinguir as florestas aleatórias quanto ao seu uso para classificação ou para regressão: quando utilizada para classificação uma RF obtém um conto de cada classe de cada árvore, em seguida classificando o voto pela maioria ocorrências. Em contrapartida, quando as florestas aleatórias são usadas para regressão, as previsões de cada árvore em um ponto são calculadas as médias.

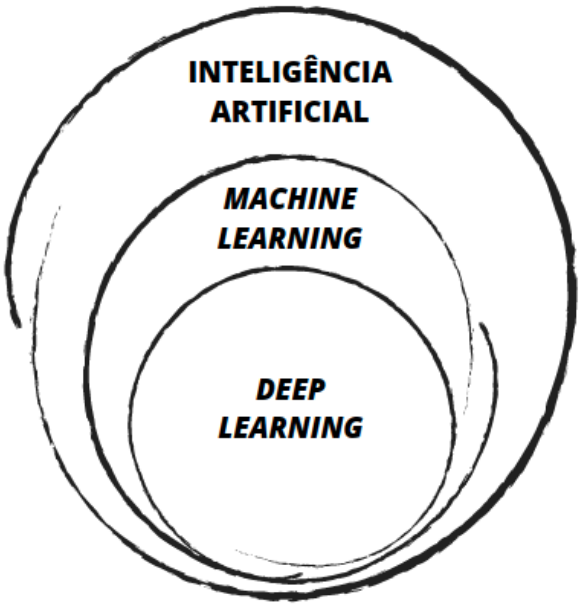
Para ser preciso, cada arvore é classificada baseada em seus valores, assim surgindo "votos" para adição em cada grupo de resultados. Em um estudo, as florestas aleatórias foram utilizadas para capturar os melhores indicadores que a possam afetar a saúde (balanços) de uma instituição financeira, podendo alertar as falhas (PETROPOULOS *et al.*, 2020). Destaca-se que conforme a floresta aumenta seu número de árvores, a capacidade preditiva do método aumenta, conforme a correlação entre as árvores diminui. Assim um número considerável de preditores pode fornecer maior capacidade de generalização, que é o caso do modelo apresentado pelo autor. Ainda, o desempenho da solução depende do número de parâmetros a serem utilizados para cada divisão de um nó (m), uma vez que (m) sendo baixo pode diminuir o peso de cada árvore na floresta. Assim se faz necessário, para um desempenho aceitável do método definir um valor ideal de (m), por meio de um técnicas de ajustes que buscam os melhores valores de parâmetro. No caso, utilizou-se a abordagem de validação cruzada (em inglês *cross-validation*), onde a grade de pesquisa abordada é bidimensional, tendo uma gama de valores para o número de divisões (m) em cada árvore para o número de árvores a serem geradas.

2.4 *DEEP LEARNING*

No contexto geral de inteligência artificial é comum ver os termos *machine learning* (aprendizagem de máquina) e *deep learning* (aprendizado profundo) serem confundidos. No entanto, para melhor compreender este trabalho de conclusão, se faz necessário diferenciar tais conceitos. Em uma visão simplista, a Figura 10 apresenta como os principais elementos (e termos) da inteligência artificial se relacionam, percebendo-se que a técnica de aprendizado profundo é uma subárea de *machine learning*.

Para compreender as diferenças, portanto, considera-se um experimento que tem por objetivo de reconhecer imagens de notas de US$ 1.00. Para isso, os modelos preditivos serão implementados em um *dataset* com milhares de fotos de dinheiro, dentre elas as notas de 1 dólar. De maneira geral, a abordagem de *machine learning* não pode analisar as fotos em si; mas sim, as informações que estão nas imagens devem ser classificadas, por meio da aprendizagem supervisionada (TAULLI, 2020). No entanto, mesmo que este treinamento tenha extraído resultados satisfatórios, pode-se sugerir que, ao invés de avaliar as imagens diretamente, sejam avaliados os *pixels* de cada foto para encontrar os padrões. Para fazer isso, provavelmente, será realizado um processo para capturar características específicas padrões das notas (valor, marca d'água, desenho de George Washington, grande selo etc), refinando capacidade preditiva

Figura 10 – Níveis dos principais componentes da Inteligência Artificial



TAULLI (2020)

do modelo. Mesmo assim, podem existir muitas variações nas fotos: amassados, rasgos, sujeira, pigmentação, notas falsificadas e afins. Com *deep learning*, no entanto, é possível resolver esses outros problemas (TAULLI, 2020). Uma vez que algumas características não são rotuladas (classificadas), sua abordagem parte de uma análise pixel a pixel, e na sequência encontra relações utilizando redes neurais. Esse tipo de sistema permite o processamento de dados para encontrar relacionamentos e padrões muitas vezes não vistos por seres humanos (TAULLI, 2020). A palavra *deep* (profundo) se refere ao número de camadas ocultas nas rede neurais as quais realizam os cálculos e fornecem grande parte do poder de aprendizagem. Aplicando redes neurais, a aprendizagem ocorre com o fortalecimento ou enfraquecimento de cada conexão neurônio a partir de seus pesos atribuídos.

Por ser uma área recente, o *Deep learning* possui vários desafios, tendo ainda aplicações restritas. Soma-se também que suas técnicas tem melhor desempenho quando os resultados podem ser quantificados, TAULLI (2020). O autor destaca ainda outros desafios dos algoritmos de aprendizado profundo que também devem ser considerados: (1) caixa preta, onde os modelos da técnica de predição facilmente possuí milhões de parâmetros que envolvem as camadas ocultas, sendo assim, a complexidade de interpretação e organização do modelo preditivo podem ser além da capacidade humana; (2) dados, que por sua vez o desafio principal é de processar enormes volumes de informações para reconhece-los, exigindo *datasets* que poucas (e grandes) organizações possuem acesso; (3) ideias de senso comum, pensamento conceitual e estruturas hierárquicas, onde a tecnologia pode se confundir na sua entrega de resultados por falta de informações e regras claras; (4) causalidade, uma vez que o objetivo é encontrar correlações; e (5) recursos, pois envolve uma grande exigência de poder computacional. Todavia, apesar dos

grandes obstáculos, a área possui recorrentes aumentos nos investimentos, servindo de incentivo para as áreas de computação e engenharias para desenvolver novas abordagens de aplicação.

2.5 PRINCIPAIS TÉCNICAS DE *DEEP LEARNING*

O recente interesse nas áreas de *deep learning* (aprendizado profundo) tem despertado resultado em uma crescente contribuição acadêmica nas áreas das engenharias e computação. Algumas contribuições na área médica permitem que em um procedimento de radiologia as varreduras de imagens sejam aceleradas, ao mesmo tempo que se aumenta o resultado dos exames e a lucratividade da instituição, uma vez que que as técnicas de aprendizado profundo são aplicados aos *scanners* que realizam os exames, (MEDICAL, 2018). Outra abordagem que tem auxiliado médicos e hospitais são estudos envolvendo *deep learning* detectando o mal de Alzheimer. Nesse contexto, o diagnóstico precoce da doença é fundamental, onde as ferramentas de aprendizado profundo em alguns estudos tem encontrado precisão nos seus diagnósticos de 92% a 98% (TAULLI, 2020). Enfim, pesquisadores tem apresentado diversas técnicas e áreas que se aplicam ao problema de aprendizado profundo. Propõe-se, nesta seção, uma breve descrição das técnicas mais comuns conforme o estudo investigativo de revisão sistemática produzido em 2.6.

2.5.1 Redes Neurais Convolucionais

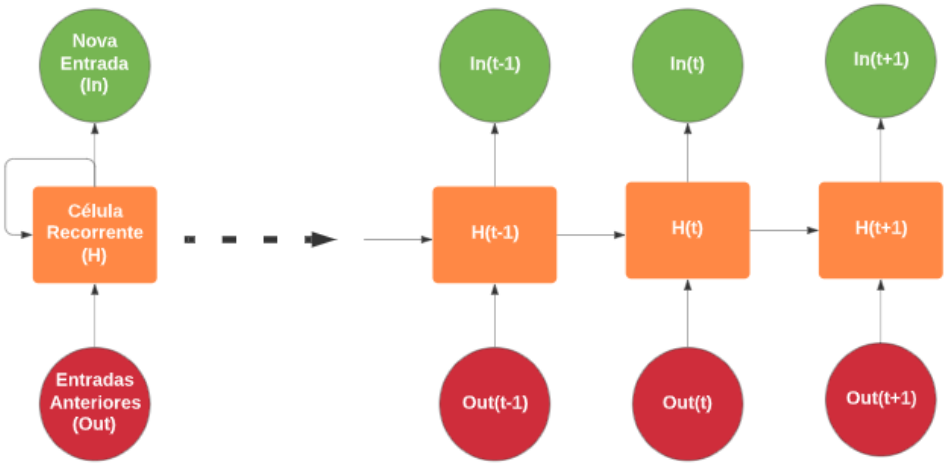
Uma rede neural convolucional ou CNN consiste em modelo preditivo amplamente utilizado no campo de visão computacional que busca captar imagens nas suas entradas, e da mesma forma que as redes neurais, atribuir importância (pesos e vieses) a vários aspectos e ser capaz de diferenciar um nodo do outro. Assim, as redes convolucionais se caraterizam por uma grande semelhança as ideias das redes neurais: retro-propagação e funções de ativações não lineares. No entanto, ao contrário das técnicas de *machine learning* supervisionadas, onde as características devem ser rotuladas pela intervenção humana, os modelos de redes neurais convolucionais tem a capacidade de aprender essas classificações de características. Ainda, da mesma forma que as redes neurais, a conectividade das unidades (neurônios) são inspiradas no cérebro humano. Tais unidades respondem individualmente a determinados estímulos conforme cada nível de convolução. Um dos seus diferenciais em relação às redes neurais convencionais é que usando pode-se obter melhores resultados a partir do conjunto de dados envolvido (ACADEMY, 2019). As redes neurais convolucionais se diferenciam a medida que se reduz o número de valores e à capacidade de reutilização dos pesos de cada nodo, assim sendo refinada para entender melhor a sofisticação da imagem. Por exemplo, supondo-se um cenário em que se necessita realizar um processo de reconhecimento de imagens, é possível imaginar o quão complexo seria uma relação entre nodos de uma rede neural. Haveriam outras complicações como o próprio *overfitting*. Para lidar com esses problemas, uma solução é usar as redes neurais convolucionais. Na aplicação de reconhecimento de imagens, o sistema pega uma imagem e a processa em várias fases

(convolução). Por exemplo: o nível inicial é responsável pela detecção de ângulos, o segundo linhas; o terceiro formas; e finalizando identificando os objetos. Autores destacam estudos para prever falência de instituições e empresas utilizando técnicas de redes neurais convolucionais.

2.5.2 Memória Longa de Curto Prazo

Para que se possa compreender as memórias longas de curto prazo, em inglês *Long Short-Term Memory* (LSTM), se faz necessário entender como funcionam os modelos preditivos de redes neurais recorrentes e os problemas que podem ocorrer, como os problemas das memórias de curto prazo. As redes neurais recorrentes, além de processar os dados na entrada atual, processam as entradas anteriores ao longo do tempo. Para exemplificar, pode-se observar esses modelos preditivos em situações quando se digita mensagens em aplicativos de mensagens nos *smartphones*. À medida que ocorrem as digitações, o sistema tenta prever às próximas palavras. As redes neurais recorrentes são essencialmente uma sequência de redes neurais (vistas nas seções anteriores) que reintegram novas informações umas das outras nas suas entradas conforme representado na Figura 11. No diagrama, cada célula recorrente (laranja) recebe novas entradas In (verde) e o estado do instante anterior, gerando, assim, uma saída Out (vermelho).

Figura 11 – Funcionamento de uma rede neural recorrente



TAULLI (2020)

No entanto, se uma sequência de informações de um *dataset* for muito longa, o modelo preditivo terá dificuldades para transportar os dados de um nível para outro. Assim, para uma análise de texto utilizando redes neurais recorrentes, pode ser que algumas partes relevantes sejam descartadas, perdendo informações que podem alterar o resultado e interpretação. Durante a etapa de retro-propagação, o sistema pode sofrer com o problema da dissipação do gradiente[10]. Esse problema consiste na dissipação do gradiente, quando ele decai à medida em que

[10] Gradiente: gradientes são os valores usados para atualizar os pesos de cada neurônio na técnica das redes neurais

os dados se propagam dentro de cada nodo iterativamente com o passar do tempo. Se um valor de gradiente se torna extremamente pequeno, ele não irá contribuir com o aprendizado. Assim, nas redes neurais recorrentes, as camadas que recebem uma pequena atualização gradiente param de aprender, deixando os neurônios inativos, podendo "esquecer"o que foi repassado nas sequências mais longas, recebendo o nome de memória de curto prazo.

Um dos algoritmos que tendem a resolver esses problemas, são as memórias longas de curto prazo. Conforme definem os pesquisadores PAL; GHOSH; NAG (2018), o LSTM possui unidades de memória na camada oculta recorrente, que contém células de memória que armazenam o estado temporal da rede. Essa rede possui unidades multiplicativas chamadas de portas (ou portões) para controlar o fluxo de informações na rede. Cada unidade de memória na arquitetura original possui uma porta de entrada e uma porta de saída. A porta de entrada controla o fluxo de ativações de entrada na célula de memória e a porta de saída controla o fluxo de saída de ativações de célula no resto da rede. Uma porta de esquecimento é adicionada ao bloco de memória que dimensiona o estado interno da célula antes de adicioná-lo como entrada à célula por meio da conexão autor-recorrente da célula, ou seja, redefinindo a memória da célula. Além disso, o LSTM também pode conter conexões que ajudam a aprender o tempo preciso das saídas. Em suma as memórias longas de curto prazo possuem mecanismos internos chamados de portões que regulam o fluxo das informações. Tais portões aprendem os dados em que uma sequência são relevantes para mantê-los ou descartá-los, em outras palavras, são capazes de aprender diferentes conexões ao longo do tempo. Essas estruturas aplicam funções (como a função *sigmóide*) que regulam qual o percentual de cada valor deve passar pelos portões *gates*.

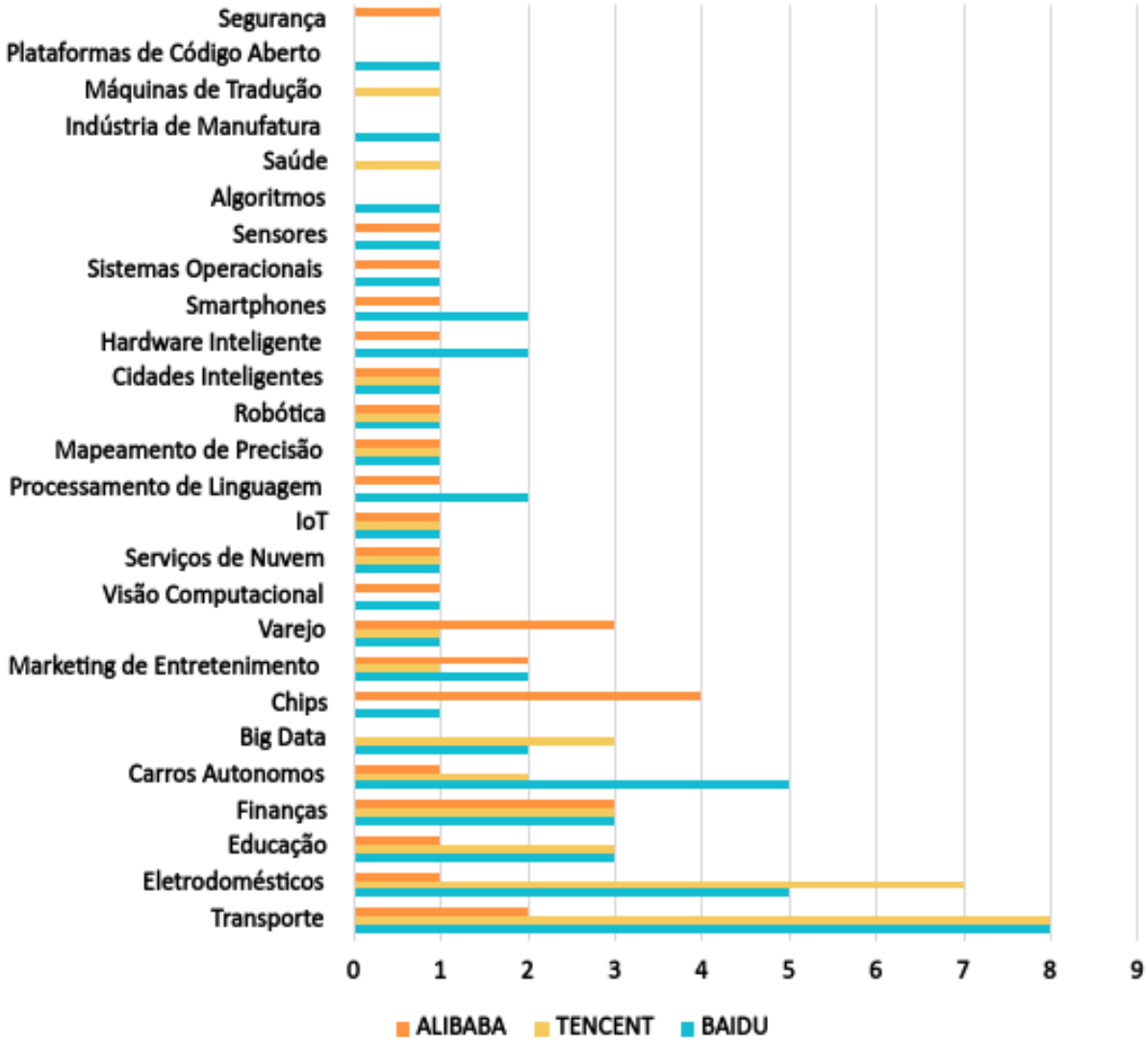
Fundamentalmente conforme a publicação do autor PHI (2018), as memórias longas de curto prazo possuem 4 portões diferentes: (1) *forget gate*, que toma as decisões de quais partes das células continuam relevantes; (2) *input gate*, que por sua vez passa o estado oculto anterior e a entrada atual para uma função, decidindo quais valores serão atualizados; e (3) *output gate*, que decide quais partes são relevantes no instante atual para gerar a saída. O procedimento permite transmitir os atributos relevantes a diante de uma extensa cadeia de dados para realizar as previsões de forma mais assertiva, ACADEMY (2019). Para entender o funcionamento deste método preditivo, considera-se um experimento que busca prever as próximas palavras de um texto. Na frase "(...) eu trabalho com inteligência artificial com meu professor, (?) ", ao ser digitado a palavra "professor"os passos dos portões podem ser traduzidos em: (1), esquecer o sujeito "eu"(*forget gate*); (2), lembrar o sujeito "meu professor"(*input gate*); e (3), utilizar a célula de estado para lembrar com o que o professor trabalha (*output gate*).

2.6 MODELOS PREDITIVOS APLICADOS NA ÁREA FINANCEIRA

Os modelos preditivos podem ser amplamente estudados em várias áreas do conhecimento, tanto humanas quanto exatas. Nas áreas da economia, modelos preditivos são necessários para a tomada de decisões políticas, investimentos e gestão de risco. Nesse contexto, a

China se colocou no lugar de uma das maiores potências do mundo quando o assunto é tecnologia. O país, que é um dos centros econômicos de desenvolvimento mais relevantes e inovadores do globo, é alvo de grandes investimentos na área de inteligência artificial. De acordo com a publicação do MIT (HAO, 2019), empresas como *Baidu*, *Alibaba* e *Tencent* que se equivalem às gigantes americanas *Google*, *Amazon* e *Microsoft* investem em diversas *startups* chinesas que englobam soluções desde cidades inteligentes até educação, conforme ilustrado na Figura 12.

Figura 12 – Investimentos das empresas Chinesas em Inteligência Artificial



TAULLI (2020)

No entanto, o crescimento acelerado de investimentos e desenvolvimento trouxe novos desafios: a China saltou da indústria de cartões de crédito diretamente para pagamentos por aplicativos *online*. Empresas como *WeChat* e a *Alipay* permitem que o usuário realize compras diretamente da sua conta corrente, mas seus serviços não permitem gastos acima dos recursos reais, enquanto se espera pelo próximo salário. Para solucionar esse problema, a empresa chinesa de inteligência artificial *Smart Finance* (LEE, 2019), depende exclusivamente de algorit-

mos para realizar pequenos empréstimos. Suas implementações preditivas não analisam apenas métricas que um analista de crédito de banco conseguiria avaliar, mas também possui o poder de previsão a partir de um conjunto de dados que parecem ser irrelevantes para os agentes de crédito convencionais. Por exemplo, a velocidade com a qual o tomador de empréstimo digitou a suas informações, a quantidade percentual de bateria que resta no celular e o dia ao qual foi solicitado o dinheiro extra. Esses dados, que passam despercebidos por bancos e instituições de crédito, podem auxiliar o financiador a tomar decisões mais assertivas. A *Smart Finance* utiliza milhares de características fracas que estão relacionadas com a credibilidade do cliente, a partir do treinamento de dados com seus modelos preditivos. Ainda, os novos cadastros de novos clientes ampliam a precisão dos algoritmos, possibilitando que a empresa estenda suas áreas de atuação a grupos ignorados pelo setor bancário Chinês, como jovens e imigrantes. No final de 2017 a empresa de *Ke Jiao* realizou cerca de 2 milhões de empréstimos por mês com taxas de inadimplência inferiores aos bancos tradicionais utilizando apenas algoritmos de aprendizado de máquina. Muitas destas técnicas utilizadas pela empresa chinesa partem dos conceitos de *machine learning*.

Prever o mercado de ações, a solvência de bancos e instituições, a taxa de risco de investimento no ramo imobiliário, crescimento dos indicadores econômicos são de extrema importância para governos, entidades de desenvolvimento, empreendedores e *holdings* para a tomada de decisão de investimento. Todo risco que investidores e empreendedores assumem é calculado de acordo com a sua expectativa de um retorno maior futuro (MISES, 2017). Sendo assim, as ferramentas de previsão podem facilitar na tomada de decisão uma vez que essas avaliações são um desafio complexo que possuem um grande número de variáveis para medir o momento mais confortável de entrada ou saída no mercado. Na realidade, finanças é uma das áreas mais estudadas com aprendizado de máquina há já 40 anos (OZBAYOGLU; GUDELEK; SEZER, 2020). As técnicas de *machine learning* e *deep learning* são as ferramentas que os pesquisadores utilizam para desenvolver modelos que podem fornecer soluções de trabalho em tempo real para o setor financeiro. Há uma quantidade imensa de aplicações de métodos preditivos para várias subdivisões de interesse no mercado financeiro que podem ser encontradas na literatura, existindo oportunidades em várias áreas específicas. Para tanto, foram tabuladas as características representativas dos estudos relevantes na revisão sistemática, a fim de fornecer o máximo de informações possíveis nas revisões realizadas. Pode-se notar também que algumas sobreposições entre diferentes implementos e termos para alguns artigos. Houve três razões principais para isso: em alguns trabalhos, vários problemas foram tratados separadamente; em outros, foram propostas novas técnicas de desenvolvimento particulares; enquanto em outros casos, o papel pode caber diretamente em várias áreas de implementação devido à estrutura de pesquisa. Após, agrupa-se em uma breve descrição de implementação de algoritmos preditivos para a área investigada, tais como: algoritmos de negociação; detecção de fraudes; gestão de portfólio; mineração de texto; e por fim, avaliação de risco.

2.6.1 Revisão Sistemática

Para desenvolver este projeto, iniciou-se com o processo de revisão sistemática, (SAMPAIO; MANCINI, 2007). Seguindo processo do método de revisão sistemática, deve-se iniciar definindo um conjunto de termos ou palavras chaves a serem utilizadas para pesquisa bibliográfica. Como palavras chaves utilizou-se as seguintes: "*economic*", "*analysis*", "*crisis*", "*prediction*", "*machine learning*"e "*algorithm*". A fonte de pesquisa foi o portal *ScienceDirect*[11]. A combinação dessas palavras, utilizadas para filtrar os documentos, foi integrada a outros filtros disponíveis no portal (ano da publicação e local). Foram selecionados trabalhos dos anos de 2018 a 2021, publicados nos títulos *Applied Soft Computing*, *Expert Systems with Applications*, *Journal of Cleaner Production*, *International Journal of Forecasting*, *Technological Forecasting and Social Change* e *Procedia Computer Science*. Tais revistas foram selecionadas por publicarem trabalhos nas áreas de interesse desta pesquisa, delimitando assim o universo da busca.

Aplicados os filtros, obteve-se 115 trabalhos, que foram selecionados de acordo com os seus títulos, tendo-se excluído 50 publicações consideradas fora do contexto da pesquisa. Em seguida, foi feita a leitura dos resumos. A partir da leitura, identificou-se que apenas 30 trabalhos estavam relacionados ao tema da pesquisa. Ainda, a partir das leituras das contribuições selecionadas, destacou-se 14 artigos sendo os mais relevantes. Em um análise mais aprofundada, detectou-se que 3 deles tratavam do tema da Econometria, sendo descartados. Considerou-se, assim, para esse processo de análise sistemática, um corpus final de 11 artigos científicos. A partir do aprofundamento, identificou-se que os principais elementos a serem analisados são os seguintes: método, algoritmos (Tabela auxiliar 1 organiza suas abreviações), ferramentas, ano de publicação e resultados.

Esses elementos foram extraídos dos artigos lidos, na forma de dados, que foram tabulados e são apresentados na Tabela 2. Para melhor entendimento e interpretação, tabelas auxiliares foram adicionadas a tabela principal, possibilitando um maior detalhamento dos trabalhos. Assim, as Tabelas: 3 apresenta os problemas pesquisados e suas metodologias de estudo; 4 e 5 respectivamente mostram os períodos dos conjuntos de dados e suas descrições (instituições, atributos, regras); 6 e 7 abordam variáveis de medição de resultados e os resultados quantificados em si, nesta ordem.

Além da revisão sistemática, outras contribuições foram consideradas relevantes e necessárias na construção deste manuscrito. Destacam-se os seguintes: (FISCHER; KRAUSS, 2018; BLUWSTEIN *et al.*, 2020).

No primeiro artigo são implementadas redes LSTM para prever os retornos diários do índice S&P 500[12]. Seu conjunto de dados avaliados variam no período de 1992 até 2015, onde seus atributos principais de avaliação são consolidar listas em uma matriz binária, indicando se

[11] https://www.sciencedirect.com/

[12] S&P 500: índice das 500 maiores empresas americanas listadas em bolsa de Nova Iorque.

Tabela 1 – Abreviações das nomenclaturas dos métodos preditivos

Método	**Sigla**	**Método**	**Sigla**
Regressão Logística	LogR	Árvores de Decisão	CART
Florestas Aleatórias	RF	Máquinas de Suporte Vetorial	SVM
Redes Neurais	NN	Aumento de Gradiente Extremo	XGBoost
Rede *feed forward* de DP [a]	MXNET	Combinação de Previsão	CLS
Suporte de Regressão Vetorial	SVR	Rede Elástica	EL
Árvores Extremamente Aleatórias	ERT	Rede Neural Profunda	DNN
Memória Longa de Curto Prazo	LSTM	Análise Discriminante Linear	LDA
RF de Inferência Condicional	CRF	Algoritmo C.45	C45
Classificador de Bayes Naive	NB	K-Vizinhos Mais Próximos	KNN
Classificador Essemble	CE	Subespaço Aleatório	RS
RS2_ER [b]	RS2	Análise Discriminante Multivariável	MDA
Modelo de Gráfico Generativo	DBN	Rede Neural Convolucional	CNN
Máquina Boltzmann Restrita	RBN	Perceptron Multicamadas	MLP
Stacked Autoencoder	SAE	Programação Genética	GP
Rede Bayesiana Discreta	BND	Rede Bayesina Com Variável Latente	BNVL

Fonte: O Autor (2021)

[a] DP: *Deep Learning*
[b] RS2_ER: Variação de método construído proposto de subespaço aleatório (RS), (WANG *et al.*, 2020)

a ação é um constituinte do índice nos meses subsequentes. A metodologia utilizada consiste, em 5 etapas: (1) dividir os dados sem tratamento em períodos de estudo; (2) discutir o espaço de recursos e as metas necessárias para o treinamento e fazer previsões; (3) e (4) abordar a aplicação das redes LSTM e revisar outros métodos preditivos (LogR, RF e DNN); (5) desenvolver uma abordagem comercial. Os resultados são discutidos em 3 etapas. Primeiramente, analisa-se os retornos de antes/depois dos custos das transações para constatar o desempenho dos modelos (LSTM *versus* RF; LSTM *versus* DNN; *versus* LogR). Em segundo lugar, deriva-se alguns padrões reconhecidos, revelando assim fontes de lucro. Por fim, desenvolve-se uma estratégia de negociação simplificada com base nas descobertas. Utilizando LSTM pode-se capturar padrões mais visíveis com as regras estabelecidas de negociação. Os resultados encontrados pelo método preditivo, seguindo sua metodologia de estudo, o LSTM apresentou-se superior em relação a outras abordagens em quase todos os contextos do estudo.

No segundo estudo são desenvolvidos modelos de previsão de crises financeiras usando aprendizado de máquina com dados de 17 países datados de 1870 até 2016. Os atributos utilizados nos preditores mais relevantes são: o crescimento do crédito e a inclinação da curva de juros, tanto doméstica quanto globalmente. O desempenho dos seguintes métodos preditivos foram avaliados: CART, RF, ERT, SVM e NN. Na análise principal, a validação cruzada é utilizada para avaliar a previsão fora da amostra desempenho dos modelos. Isso envolve calibrar os modelos explorando todos os dados em seleções aleatórias de 80% desses dados, chamados de conjunto de treinamento, e avaliação dos outros 20% restantes das observações, o conjunto de

Tabela 2 – Resultados da Área Científica

Referência	**Problema/Metodologia**	**Conjunto de dados**	**Resultados**
(CHATZIS *et al.*, 2018)	1. - na tabela 3. Algoritmos: LogR, CART, RF, SVM, NN, XBoost, MXNET e CLS	1. - nas tabelas 4 e 5	1. - nas tabelas 6 e 7. Melhor resultado: MXNET.
(GHOSH; JANA; SANYAL, 2019)	2. Algoritmos: SVR, EN, RF, ERT, Boosting, DNN, e LSTM	2	2. Melhores resultados: DNN e Boosting.
(PETROPOULOS *et al.*, 2020)	3. Algoritmos: LogR, LDA, RF, SVM, NN e CRF	3	3. Melhor resultado: RF.
(CHOI; SON; KIM, 2018)	4. Algoritmos: SVM, NN, C4.5, NB, LogR, KNN e CE	4	4. Melhor resultado: CE.
(OZBAYOGLU; GUDELEK; SEZER, 2020)	5. Algoritmos: DBN, RBN, SVM, MLP, GP, CNN, LSTM, LogR, RF, XBoost	5	5. Melhor resultado: DBN.
(GOGAS; PAPADIMITRIOU; AGRAPETIDOU, 2018)	6. Algoritmos: SVM	6	6. O SVM exibe uma forma consistente nos resultados.
(WANG; ZONG; MA, 2020)	7. Algoritmos: LSTM, NN, SVR	7	7. Melhor resultado: LSTM.
(WANG *et al.*, 2020)	8. Algoritmos: SVM, AdaBoost, Ensacamento, RS, RS2	8	8. Melhor resultado: modelos RS.
(MASMOUDI; ABID; MASMOUDI, 2019)	9. Algoritmos: BND, CART, SVM, BNVL	9	9. Melhores resultados: BND e BNVL.
(HUANG; YEN, 2019)	10. Algoritmos: SVM, HACT, GA, XBoost, GPC, RF, TSE, LC, BND, DNN	10	10. Melhor resultado: XBoost.
(QU *et al.*, 2019)	11. Algoritmos: MDA, LogR, NN, SVM, DNN e CNN	11	11. Melhores resultados nos métodos preditivos vem da diversificação da quantidade de fontes de dados.

Fonte: O Autor (2021)

teste. Todos os modelos identificam consistentemente preditores semelhantes para crises financeiras, embora existam algumas variações ao longo do tempo, refletindo mudanças na natureza da política monetária global e sistema financeiro nos últimos 150 anos. No estudo, o modelo que resultou no melhor desempenho preditivo, ERT, corresponde em identificar uma crise com uma probabilidade de 9,6%.

A partir do processo de revisão sistemática identificou-se domínios de aplicação dos métodos preditivos à área financeira. Os princípios domínios são descritos na seção seguinte.

Tabela 3 – Problema e Metodologia Abordados

Número de Referência	Descrição
1	Previsão de quebra da bolsa de valores, onde se compara algoritmos para previsão de quebra da bolsa para em dados de 1 e 20 dias.
2	Estimar o valor absoluto de preços de ações ou retornos futuros. Calcula-se a série de retorno dos índices de ações, usando resultado do teste para a probabilidade de uma condição existir.
3	Capturar variáveis determinantes que afetam a saúde de uma instituição financeira e paralelamente, desenvolver um sistema de alerta para prever falhas. Utilizar a base de dados para prever insolvência de bancos dos EUA.
4	Prever e avaliar dificuldades financeiras das fases iniciais de um contratante em um projeto de construção em 2 a 3 anos antes. Treinar vários modelos de predição cujas decisões são combinadas.
5	Avaliação de risco; Identificar o risco atribuído a um ativo, instituição ou investidor. Previsão de falência, fracasso empresarial e risco de crédito coorporativo. Comparar vários métodos de vários trabalhos. Cada trabalho comparado possui um *dataset específico*.
6	Metodologia baseada em SVM para prever falências de bancos e instituições financeiras. Separar dados em duas classes, para encontrar um hiperplano que maximiza a distância entre dois pontos.
7	Desenvolver um classificador prórpio e comparar a outros modelos preditivos para prever início de crise. Propor uma variação de LSTM e comparar o seu resultado com NN e SVR.
8	Propor uma nova estrutura de precisão e previsão para crises financeiras, utilizando dados financeiros e não financeiros (não textuais).
9	Avaliar o risco de crédito, permitindo uma análise mais profunda análise de inadimplência de clientes. Comparar o desempenho do BNVL com os outros métodos.
10	Revisão de métodos preditivos para prever dificuldades financeiras. Comparar o desempenho de cada modelo para o conjunto de dados.
11	Revisar Modelos de previsão de falências de instituições financeiras. Discussão e comparação dos modelos apresentados.

Fonte: O Autor (2021)

2.6.2 Áreas de Aplicação Identificadas

A revisão sistemática possibilitou a identificação e o reconhecimento de áreas prioritárias onde os métodos preditos vem sendo aplicados, nomeadamente: algoritmos de negociação, detecção de fraudes, gestão de portfólio, mineração de texto e avaliação de risco.

Os algoritmos de negociação são definidos como robôs de decisão de compra e venda

Tabela 4 – Detalhamento de datas do conjunto de dados dos *datasets*

Número de Referência	Data	Número de Referência	Data
1	2011-2017	7	2008-2018
2	2012-2017	8	Não Informado
3	2008-2014	9	1990-2012
4	2011-2017	10	2010-2016
5	Não Informado	11	Não Informado
6	2007-2013		

Fonte: O Autor (2021)

de ativos. Essas decisões são realizadas a partir de regras simples, modelos matemáticos ou no caso dos aprendizados de máquina ou aprendizado profundo, funções complexas para atingir o pico e o vale de uma ação. Na maioria dos casos de aplicação das técnicas de algoritmos de negociação, são combinados modelos de previsão com os preços do mercado em tempo real. Como resultado, tais técnicas retornam indicadores de tendências futuras, (OZBAYOGLU; GUDELEK; SEZER, 2020).

No cenário de detecção de fraudes financeiras tem-se uma das áreas em que governos, autoridades e instituições privadas estão tentando encontrar uma solução permanente. Historicamente, existem vários casos de fraudes financeiras, tais como: fraude de cartão de crédito, lavagem de dinheiro, fraude ao consumidor, evasão fiscal, e fraude bancária. Essas fraudes são notícias cotidianas em todo o mundo. Observou-se que os estudos neste contexto são principalmente sobre detecção de anomalias, sendo tratados portanto, como problemas de classificação.

A gestão de um portfólio é a prática de escolha de vários investimentos para uma determinada pessoa ou empresa. Como visto nas outras aplicações financeiras, versões ligeiramente diferentes desse problema são desafios cotidianos, uma vez que o mercado varia a cada instante, como novos eventos e decisões humanas. Em geral, essa aplicação cobre as seguintes áreas estreitamente relacionadas: otimização de portfólio, seleção de ativos, alocação de portfólio. Assim, alguns autores classificam o gerenciamento de portfólio como um problema de otimização, identificando o melhor movimento possível para selecionar os ativos de melhor desempenho em um determinado período. Como resultado, há modelos de algoritmos evolutivos[13] que são desenvolvidos para essa finalidade, embora outros pesquisadores desenvolveram modelos de aprendizagem de máquina e obtiveram desempenhos superiores, (OZBAYOGLU; GUDELEK; SEZER, 2020).

Quanto a mineração de texto, percebe-se que, com a rápida disseminação das mídias sociais em tempo real, a pesquisa de informações baseadas em textos tornou-se de fácil acesso para o desenvolvimento de um modelo para soluções nos mercados financeiros. Como resultado, os estudos financeiros de mineração de texto se tornaram muito populares no mundo acadêmico nos últimos anos, (OZBAYOGLU; GUDELEK; SEZER, 2020). Mesmo que alguns desses estudos

[13] Algoritmos evolutivos: técnicas inspiradas na evolução biológica, como reprodução, seleção e recombinação.

Tabela 5 – Detalhamento dos atributos presentes nos *datasets*

Número de Referência	**Descrição do *dataset***
1	Cotações médias anuais de preço de índice de ações, títulos de 10 anos do governo, taxa de câmbio (*versus* dolar), preço do ouro, VIX, preço de *commodity*.
2	Preços de fechamento diários de bolsas de valores coletados no repositório Metastock.
3	Informações de quebras e falhas de instituições financeiras dos EUA do banco de dados *Federal Deposit Insurance Corporation* (FDIC). Cada banco, é categorizado como solvente ou insolvente com base nos indicadores do *dataset*.
4	Descrição Completa na tabela principal.
5	Descrição Completa na tabela principal.
6	Informações de todos os bancos e instituições que estavam ativadas e faliram nos EUA.
7	Utilizando o *Shanghai Stock Exchange Composite* (SSEC) índice que reflete a oscilação do mercado de ações chinês. O preço de fechamento da bolsa, retorno logarítmico e volatilidade realizado índice SSEC. Também, a renda fixa local, os mercados imobiliários e moeda, a macroeconomia doméstica, o sentimento do investidor e o mercado de ações dos EUA.
8	Utilizadas variáveis de índices financeiros como aspectos de solvência, lucratividade, capacidades operacionais, capacidades de desenvolvimento, capacidade de expansão de de capital. Dados divididos em 3 grupos: Recursos financeiros ($F1$), Recursos de Sentimento quanto a empresas ($F2$) e Recursos Textuais ($F3$).
9	Conjunto de dados fornecidos pelo Banco Central da Tunísia, contendo 9 variáveis que descrevem contratos de empréstimos concedidos por vários bancos tunisianos. Dentre os atributos, destaca-se informações de empréstimo e se o provedor de crédito é um banco público.
10	Dados financeiros fornecidos pelo *Taiwan Economic Journal*. O banco de dados destaca 32 empresas que encontraram dificuldades financeiras no período estudado, que apresentam dados trimestrais dos seus demonstrativos financeiros.
11	Descrição Completa na tabela principal.

Fonte: O Autor (2021)

estejam diretamente interessados na análise de sentimentos (risco e retorno), existem diversas implementações que estão direcionadas para a recuperação de conteúdo de notícias, finanças, declarações, divulgações, entre outro, através da análise do contexto do texto.

Outra abordagem que tem despertado interesse de pesquisadores é a avaliação de risco. Identificar previamente o risco que envolve um indicador ou ativo, entender o mercado, crises financeiras, ciclos econômicos, flutuações econômicas, prever o fracasso empresarial, entre ou-

tros, é um desafio centenário para economistas, burocratas e analistas. Como exemplo, a crise das hipotecas (crise imobiliária de 2018) com base na avaliação não apropriada de risco entre instituições financeiras resultou em uma grande recessão, levando a falência de um dos bancos americanos mais importantes, o *Lehman Brothers* (1850-2008), (BAILY *et al.*, 2008). Todos os elementos constituem indícios, achados e fundamentos que alimentam *datasets*, e por consequência possibilitam a criação de modelos preditivos.

Várias teorias já foram propostas para explicar as crises financeiras. No entanto, a busca por uma resposta concreta ainda é uma atividade muito desafiadora, uma vez que são fenômenos raros e complexos mas que tem um impacto gigantesco na sociedade. Alguns autores usam metodologias baseadas em máquinas de suporte vetorial (GOGAS; PAPADIMITRIOU; AGRAPETIDOU, 2018) para prever a solvência de bancos, assim podendo encontrar uma falência futura de instituições financeiras. Contribuindo para previsão de falências, outros autores revisaram o desempenho de modelos, além do das máquinas de suporte vetorial, tais como o modelo de regressão logística (QU *et al.*, 2019).

Outros trabalhos ainda, desenvolveram classificadores para identificar turbulências no mercado de ações e, posteriormente, utilizaram o seu método para alertar sobre um *crash* no mercado de ações interpretando-o como uma crise (WANG; ZONG; MA, 2020). Na mesma linha, outros trabalhos acadêmicos também criaram suas novas estruturas de precisão e previsão para crises financeiras (WANG *et al.*, 2020).

2.6.3 Métricas de Avaliação Aplicadas

A partir da literatura estudada, as métricas mais observadas para avaliação de desempenho dos modelos selecionados são: acurácia, *Recall*, precisão, *F1-score* e a curva ROC (*Receiver Operating Characteristic*). Ademais, ressalta-se que existem abordagens que buscam avaliar o desempenho a partir de métricas como a matriz de confusão. Com os dados quantificados obtidos, pode-se então, encontrar o melhor modelo para os modelos selecionados no *dataset*.

A matriz de confusão é um tabela que mostra as frequências de classificação. Tais frequências são dividas em quatro grandes grupos: verdadeiro positivo (VP), quando no conjunto real de dados foi previsto corretamente; falso positivo (FP), quando o conjunto real de dados foi previsto incorretamente; verdadeiro negativo (VN), quando o conjunto real da classificação que não está sendo procurada foi prevista corretamente; e falso negativo (FN) quando o conjunto real da classificação que não está sendo procurada foi prevista incorretamente. Para exemplificar, pode-se supor um modelo para classificar fotos de gatos em um *dataset*. Nele, existem 100 fotos de gatos e 100 fotos que não possuem gatos. Após treinar o modelo, roda-se a matriz de confusão onde se descobre que das 100 fotos de gatos, o modelo classificou 50 corretamente, enquanto das 100 fotos de não gato, ele classificou 80, conforme a figura 13. Quando o modelo prevê um caso positivo de maneira correta (previu uma imagem de um gato

corretamente), é adicionado um ponto para o primeiro quadrante (verdadeiro positivo); caso o modelo retorne que uma imagem é de um gato, mas na realidade não é, o quadrante de falso positivo receberá outro ponto em sua contagem. No entanto, o oposto também pode ocorrer. O modelo pode retornar que uma das imagens é de um gato, quando na verdade não é, aumentando a pontuação de falso negativo. Ainda, quando a imagem não é de um gato que o modelo apontou a classificação corretamente, ocorre o verdadeiro negativo.

Figura 13 – Exemplo de matriz de confusão

Fonte: O Autor (2021)

Partindo-se da matriz de confusão, alguns conceitos importantes são decorrentes para os cálculos de métricas de resultados, por exemplo, acurácia, *recall*, precisão e *F1-score*.

A acurácia mede o quão efetivo é o modelo preditivo. Em suma, retorna quanto o modelo acertou das previsões possíveis (SOUZA, 2019). Pode-se calcular a métrica a partir da razão entre o somatório das previsões corretas (verdadeiros positivos com verdadeiros negativos) sobre o somatório das previsões. Assim, a acurácia é uma medida global para avaliação do desempenho do classificador, sendo definida pela equação 2.3:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

A precisão mede a proporção de identificações positivas que foram realmente corretas. Em outras palavras, o quão assertivo o modelo trabalhou. A precisão então, realiza a relação entre a quantidade de verdadeiros positivos que o sistema classificou sobre a soma entre verdadeiros positivos (VP) e falsos positivos (FP). A precisão é definida pela seguinte equação 2.4:

$$Precisão = \frac{VP}{VP + FP} \tag{2.4}$$

O *recall* (ou sensibilidade), é utilizado para indicar a relação entre as previsões positivas realizadas (VP) e todas as previsões que são positivas, isto é a soma de verdadeiros positivos

(VP) e falsos negativos (FN). Em suma, o *recall* é uma métrica que define o valor percentual em que modelo é capaz para prever positivos (classe que deseja-se prever). Em outras palavras, mede a proporção de positivos que foram identificados corretamente. Esta métrica é definida pela Equação 2.5, onde é a razão entre verdadeiros positivos sobre a soma de verdadeiros positivos com falsos negativos:

$$Recall = \frac{VP}{VP + FN} \tag{2.5}$$

Para complementar, a medida denominada *F1-socore* corresponde ao balanço entre a precisão e *recall* do modelo. A uma média harmônica entre o *recall* e a precisão. Ela é definida pela equação 2.6:

$$\textit{F1-score} = 2 \cdot \frac{precisão * recall}{precisão + recall} \tag{2.6}$$

É importante frisar que os cálculos visam compreender o modelo sobre os dados positivos. Quando, por exemplo, a classificação possui mais de duas classes possíveis de resposta, pode-se trabalhar sobre a classe que quer prever. No entanto pode-se também olhar para cada classe separadamente. Assim, na hora de construir a matriz de confusão, considera-se a classe que se quer prever como a classe positiva, e todo o restante como negativo, assim, inferindo uma classificação binária para compreender o quão assertivo o modelo está em prever a classe que se deseja (SOUZA, 2019).

A curva ROC (*Receiver Operating Characteristics*) é outra métrica que pode ser utilizada para os modelos de classificação selecionados. A curva AUC é derivada da curva ROC. O cálculo dessas métricas funciona de forma semelhante a matriz de confusão. Pode-se plotar a sensibilidade (Verdadeiro Positivo) e especificidade (Falso Positivo). Além disso, existe a possibilidade de passar mais um parâmetro, conhecido como *Threshold* ou *T parameter*. Esse parâmetro pode ser imputado, onde geralmente é escolhido um valor de 0,5, (SOUZA, 2019).

A curva ROC mostra o quão assertivo o modelo criado pode distinguir entre duas definições (já que é utilizado para classificação). Essas duas definições, por exemplo, podem ser 0 ou 1, ou positivo e negativo, crise ou não crise, vender ou comprar entre outros. Os melhores modelos conseguem distinguir com precisão o binômio. A curva ROC possui dois parâmetros: taxa de verdadeiro positivo (VPR) que é dado pela Equação 2.7 e taxa de falso positivo (FPR) que pode ser observada na equação 2.8.

$$VPR = \frac{VP}{VP + FN} \tag{2.7}$$

$$FPR = \frac{FP}{FP + VN} \tag{2.8}$$

Por fim, pode-se complementar a curva ROC com outra métrica de resultados: área sob a curva (AUC). Esta métrica de resultados pode variar entre 0 até 1, onde o limiar entre a classe é 0,5. Um modelo cujas previsões realizadas estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1. O diferencial do AUC é que a métrica é invariante em escala, uma vez que trabalha com precisão das classificações ao invés de seus valores absolutos. Além disso, também mede a qualidade das previsões do modelo, independentemente do limiar de classificação (RODRIGUES, 2018).

2.7 CONSIDERAÇÕES FINAIS

No intuito de extrair conclusões sobre o estudo realizado são apresentadas algumas considerações que buscam tratar de pontos importantes vistos no decorrer do desenvolvimento deste capítulo. Os estudos revisados sugerem que o aprendizado profundo (*deep learning*) tem movimentado grande interesse acadêmico para pesquisas em finanças (OZBAYOGLU; GUDELEK; SEZER, 2020).

A aplicação de métodos computacionais na área financeira abrange um grande número de contribuições, em diferentes temas, como visto na revisão sistemática. As aplicações apresentadas constituem uma grande oportunidade de estudo para as áreas do conhecimento das engenharias e que o conjunto de dados deve ser confiável e com suficiente número de atributos disponíveis. Os atributos que compõe o conjunto de dados devem se revelar relevantes e necessários para um processo de aprendizado de máquina. Para se descobrir qual o conjunto mínimo necessário de atributos, os trabalhos estudados se valem de sucessivos testes e avaliações dos resultado (processo de *feature selection*).

Sumarizando o que foi observado por meio da revisão sistemática, não basta obter-se um grande volume de dados, pois deve-se garantir a qualidade dos dados, por meio de uma distribuição correta de amostras por classe e redução de erros e ruídos. Notou-se também que melhores resultados podem ser obtidos a partir da combinação de fontes diferentes de dados (QU *et al.*, 2019) para compor um conjunto completo. Outra contribuição do estudo realizado está evidenciada na possibilidade de implementação de um pré-processamento dos atributos mais relevantes. Deve-se realizar o pré-processamento para ajuste dos valores dos atributos (discretização, normalização, entre outros) e separá-los entre os mais relevantes para obter resultados melhores (BLUWSTEIN *et al.*, 2020). Pode-se também pensar no aprimoramento de testes, onde se possa realizar a comparação dos resultados obtidos com outros classificadores, construindo assim parâmetros comparativos.

A partir da revisão realizada, pode-se atestar que para fins de predição de crises financeiras pode-se aplicar métodos de classificação baseados em *deep learning*. É possível afirmar que as contribuições destes estudos indicam que as redes LSTM constituem o modelo dominante em aprendizado profundo. Isso se deve aos bons resultados obtidos e à sua estrutura bem

estabelecida para previsão de dados de séries temporais financeiras. As implementações da área financeira têm representações de dados que variam no tempo e, assim, se adaptam a problemas preditivos envolvendo séries temporais tais como os modelos financeiros. Por se ajustar ao domínios, as abordagens LSTM oferecem bons resultados. Por fim, os principais trabalhos se baseiam na matriz de confusão, gerada pelo conjunto de dados de teste, e nas métricas de avaliação dos resultados calculadas a partir dela.

Tabela 6 – Detalhamento dos atributos de resultados do trabalho

Número de Referência	Descrição das variáveis resultado
1	*Hit Rate*: Taxa de acerto
2	NSC: Coeficiente de *Nash Sutcliffe*. Varia entre infinito e 1. Quanto mais próximo a 1, melhor a acurácia. IA: ìndice de Concordância. Valores mais próximos a 1 inferem em previsões mais eficientes, enquanto valores mais próximos a 0 sugerem previsões ruins. TI: *Theil Inequality*. Quanto mais próximo a 0 for o valor, mais satisfatório é o resultado do modelo. RMSE: Erro Quadrático Médio. Valor dimensionado de 0 a 1. Quanto mais próximo de 0, mais eficiente. MAD: Desvio Médio Absoluto. Valor dimensionado de 0 a 1. Valor dimensionado de 0 a 1. Quanto mais próximo de 0, mais eficiente. MAPE: Erro de Porcentagem Média Absoluta. Valor dimensionado de 0 a 1.Valor dimensionado de 0 a 1. Quanto mais próximo de 0, mais eficiente.
3	*G-mean*: Média Geométrica. Um fraco desempenho na previsão positiva levam a valores abaixo de 50%. Quanto mais baixos os valores, melhor o desempenho. BA: Precisão Balanceada. Se o classificador tira vantagem de macrodados a precisão balanceada caíra, apresentando problemas de desempenho. YI: Índice de Youden. Quanto mais alto o valor, melhor o indicador da capacidade do algoritmo de previsão. AUC: Área sob a Curva. Valores entre 0,5 e 1 denotam resultados positivos no método testado. Valores acima de de 0,9 são considerados excelentes. LR: Razão de Verossimilhança Negativa. Razão entre a probabilidade de prever um caso negativo quando
4	AUC: Área sob a Curva: Valores entre 0,5 e 1 denotam resultados positivos no método testado. Valores acima de de 0,9 são considerados excelentes.
5	Descrição Completa na tabela principal
6	Descrição Completa na tabela principal
7	Precisão do conjunto de teste e perda de entropia cruzada binária com base em LSTM, BPNN e SVR com tamanhos de janela (área) variados. L: tamanho da área da janela.
8	AUC: Área sob a curva. Valores entre 0,5 e 1 denotam resultados positivos no método testado. Valores acima de de 0,9 são considerados excelentes. F1 e F2 já listados na subtabela "Detalhamento dos atributos presentes no dos *datasets*". F1+F2: Resultado da aplicação do método com os dois conjuntos
9	Acurácia: Proporção de itens classificados corretamente. Precisão: Fração das previsões positivas corretas. *Recall*: Fração de todas as instâncias positivas classificadas corretamente como positivas. F1: Pontuação calculada a partir da fórmula: (2*Precisão**Recall*)/(Precisão + *Recall*).
10	Acurácia de para cada conjunto subdividido em trimestres ou mais de um trimestre:1Q, 2Q, 3Q, 4Q, 1Q-2Q, 1Q-3Q e 1Q-4Q.
11	Descrição Completa na tabela principal.

Fonte: O Autor (2021)

Tabela 7 – Resultados quantitativos do trabalho

Número de Referência	Valores de resultado
1	*hit rate* para 20 dias: Logit e CART: 34%; SVM e NN: 40%; RF: 42% XGBoost: 45% MXNET: 46%. *hit rate* para 1 dia: NN: 34%; RF: 38%; Logit, CART, SVM e XGBoost: 41%; MXNET: 52%; CLS: 59%.
2	NSC, IA, TI, RMSE, MAD, MAPE.
3	*G-Mean*: 0.898 (LogR), 0.884 (LDA) 0.992 (RF), 0.897 (SVM), 0.926 (NN) e 0.914 (CRF). LR: 0.184 (LogR), 0.209 (LDA), 0.000 (RF), 0.185 (SVM), 0.125 (NN) e 0.153 (CRF). BA: 0.901 (LogR), 0.889 (LDA), 0.992 (RF), 0.901(SVM), 0.927(NN) e 0.916 (CRF). YI: 0.803(LogR), 0.777 (LDA), 0.984 (RF), 0.802 (SVM), 0.854 (NN) e 0.832 (CRF). AUC: 0.980 (LogR), 0.973 (LDA), 0.998 (RF), 0.981 (SVM), 0.981 (NN) e 0.990 (CRF).
4	Gráfico AUC 2011 (estimado): 0.86(SVM), 0.92 (NN), 0.88 (C4.5), 0.83 (NB), 0.84 (LR), 0.91 (KNN) e 0,94 (CE). Gráfico AUC 2012 (estimado): 0.86 (SVM), 0.86 (NN), 0.86 (C4.5), 0.75 (NB), 0.89 (LR), 0.89 (KNN) e 0.91 (CE).
5	DBN apresentou precisão acima de 80%.
6	Taxa de acurácia para acerto de banco solventes: 99,40%. Taxa de acurácia para acerto de banco insolventes: 97,39%. Taxa de acurácia total: 99,22%
7	Acurácia L =22: 0.915 (LSTM), 0.863 (BPNN) e 9.19 (SVR) Acurácia L = 10: 0.288 (LSTM), 0.465 (BPNN) e 0.521 (SVR) Acurácia L = 5: 0.964 (LSTM), 0.947 (BPNN) e 0.928 (SVR) Acurácia total: 0.165 (LSTM), 0.239 (BPNN) e 0.399 (SVR).
8	AUC, F1: 88.54 (SVM), 89.19 (RS) e 91.24 (RS2). AUC, F2: 55.96 (SMV), 66.89 (RS) e 66.14 (RS2). AUC, F1 + F2: 90.88 (SVM), 91.79 (RS) 93.70 (RS2).
9	Acurácia: 93.53% (BND), 87.39% (CART), 93.61% (SVM) e 94.77%(BNVL). Precisão: 83.02% (BND), 61.51% (CART), 83.25% (SVM) e 85.31% (BNVL). *Recall*: 74.87% (BND), 65.12% (CART), 75.18% (SVM) e 81.31% (BNVL). F1: 78.73 (BND), 63.26% (CART), 79.01% (SVM) e 83.26% (BNVL).
10	Acurácia 1Q: 73.4 (SVM), 67.2 (HACT), 43.8(GA), 87.5 (Xboost), 53.1 (GPC e DBN). 2Q: 70.3(SVM), 56.3 (HACT), 35.9(GA), 89.1 (Xboost), 68.8 (GPC e DBN). 3Q: 67.2 (SVM), 53.1 (HACT), 39.1 (GA), 84.4 (Xboost), 51.6 (GPC e DBN). 4Q: 62.5(SVM), 64.1 (HACT), 37.5(GA), 82.8 (Xboost), 65.6 (GPC e DBN). 1Q-2Q: 82.8 (SVM), 64.1 (HACT), 39.1(GA), 90.6 (Xboost), 68.0 (GPC e DBN). 1Q-3Q: 79.2 (SVM), 57.8 (HACT), 39.1(GA), 87.5 (Xboost), 69.8 (GPC e DBN). 1Q-4Q: 81.3 (SVM), 56.3 (HACT), 40.6(GA), 90.6 (Xboost), 71.9 (GPC e DBN).
11	Sem resultados Quantitativos

Fonte: O Autor (2021)

3 *DEEP LEARNING* PARA PREDIÇÃO DE CRISES FINANCEIRAS

O desafio de prever crises financeiras é um debate centenário e desafiador para muitos economistas, políticos e empreendedores. Esse desafio torna-se cada vez mais complexo ao longo do tempo, uma vez que a economia não é estática e sofre constantes mudanças: novas convicções são criadas, novos setores são abertos, empresas podem fechar, novas legislações podem ser estabelecidas, enfim, vários eventos podem alterar a forma com a qual é enxergado o sistema econômico. Neste capítulo, são apresentados o percurso metodológico para avaliação das aplicações dos conceitos de *machine* e *deep learning*, desde a seleção e manipulação do conjunto de dados (*dataset*) e seus atributos, até as métricas de avaliação dos modelos preditivos selecionados. Após, as conclusões observadas a partir dos resultados obtidos. Também são apresentadas as plataformas e linguagens de programação que viabilizaram o estudo para aplicação dos métodos preditivos de *machine* e *deep learning*. Por fim, os resultados obtidos são descritos e a seleção do melhor método é apresentada em comparação com os resultados da área.

3.1 PERCURSO METODOLÓGICO

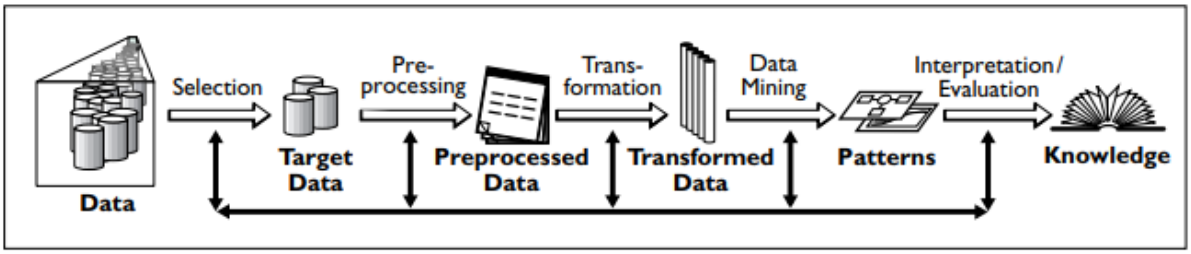
Para que se possa definir os métodos preditivos e como aplicá-los, deve-se estudar e obter as primeiras impressões sobre os cenários em que se está trabalhando. Quando estas precauções de estudo não são previamente realizadas, algumas dificuldades adicionais podem surgir posteriormente, tais como a escolha errada dos métodos preditivos, tornando a análise muito mais complexa ou inviabilizando-a. Algumas definições importantes devem ser destacadas neste trabalho: a escolha dos métodos preditivos a partir das revisões e contribuições acadêmicas, a escolha do conjunto de dados (*dataset*), a preparação deste *dataset* e da linguagem de programação. Conforme os objetivos, procura-se desenvolver um conjunto de testes a fim de identificar modelos preditivos de *machine* ou *deep learning* aptos a construírem algoritmos válidos para um conjunto de dados da área financeira.

Neste trabalho segue-se as etapas propostas pelos pesquisadores FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996), como ilustrado na figura 14, no contexto da descoberta de conhecimento em bases de dados, mas frequentemente usada em projetos de *machine learning*. O percurso previsto é composto pelas seguintes etapas:

- Etapa 1 - identificação de bases de dados (*dataset*)
- Etapa 2 - seleção de atributos
- Etapa 3 - preparação dos dados

- Etapa 4 - aplicação de algoritmos variados
- Etapa 5 - comparação dos modelos
- Etapa 6 - seleção do melhor modelo (aqueles que apresentam melhores resultados)

Figura 14 – Visão geral das etapas que constituem o processo metodológico



Fonte: FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996)

3.1.1 Etapa 1 - Identificação do *Dataset*

Crises Financeiras são eventos raros registrados na história. Ainda, são mais raros os registros e comprovantes com dados transparentes, conforme observado por BLUWSTEIN *et al.* (2020). Outro ponto a ser destacado, é que, embora existam registros de um conjunto de crises financeiras que englobam todo o planeta, como A Grande Depressão (1929) e A Crise Mundial do Subprime (2007-2008), a maior ocorrência desses eventos se apresentam principalmente em um único país ou em um pequeno grupo de países. Nesse sentido, consideram-se então a baixa frequência das crises financeiras, e portanto, são necessários explorar conjuntos de dados mais longos de séries temporais entre países disponíveis. Para a escolha da base de dados, assume-se critérios conforme conclusões realizadas na revisão sistemática: diversificação da quantidade de fontes de dados segundo os autores QU *et al.* (2019).

Para aplicação empírica, identifica-se o banco de dados de Macrohistória Jordà-Schularick-Taylor[1] disponibilizado por JORDÀ; SCHULARICK; TAYLOR (2017), que contém atributos macroeconômicos e financeiros anuais de 17 países desenvolvidos datados nos períodos de 1870 até 2016 ao qual se propõe utilizar. Para as primeiras impressões, observa-se que em cada uma das instâncias país-ano, o *dataset* possui uma variável binária que indica se o país passou por uma crise financeira (ou não) no determinado ano. Alguns autores definem as crises financeiras como "eventos durante os quais o setor bancário de um país experimenta corridas aos bancos, aumentos bruscos nas taxas de inadimplência acompanhadas de grandes perdas de capital que resultam em intervenção pública, falência ou fusão forçada de instituições financeiras", BLUWSTEIN *et al.* (2020).

[1] Banco de Dados de Macrohistória Jordà-Schularick-Taylor pode ser acessado no endereço eletrônico: http://www.macrohistory.net/

3.1.2 Etapa 2 - Seleção de Atributos

Nesta etapa, inclui-se a seleção de um conjunto de dados e concentra-se em um subconjunto de atributos (colunas do *dataset*) e amostras de dados nas quais a descoberta deve ser realizada. Posteriormente, a partir de atributos do *dataset*, visto que deseja-se prever crises com antecedência, o atributo resposta do *dataset* selecionado apresenta uma variável binária de acordo com a instância país-ano (crise ou não crise).

Diante disso, segundo BLUWSTEIN *et al.* (2020), para o conjunto de dados selecionado, os atributos com as maiores participações preditivas nos resultados são a inclinação da curva de rendimento global e crescimento do crédito global. Tais atributos são obtidos a partir de um pré processamento do *dataset* apresentado no próximo capítulo. Destaca-se também a importância das variáveis ao longo do tempo: os sistemas financeiros sofrem constantes mudanças, onde durante uma série de crises financeiras (não globais) ocorridas na década de 1990, o crédito interno de cada país e o coeficiente de dívida interna são os principais indicadores. No entanto, em uma crise financeira global, o indicador mais importante é a expansão crédito global.

3.1.3 Etapa 3 - Preparação dos Dados

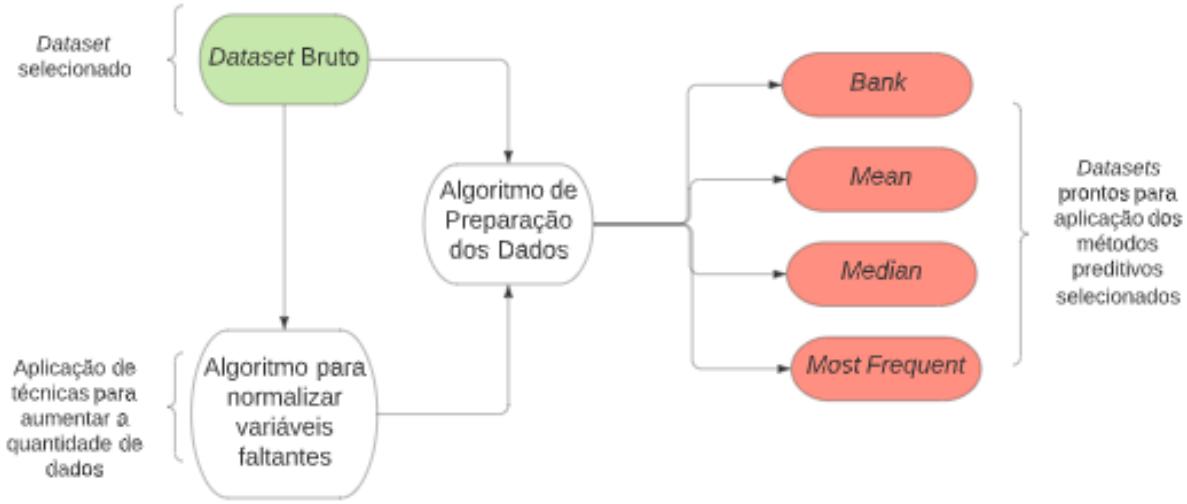
Na etapa 3, o objetivo é realizar operações básicas, como remover *outliers* e ajustar informações necessárias para modelar e/ou decidir sobre quais estratégias selecionar para lidar com a falta de dados em algumas células. Busca-se então, encontrar recursos úteis para representá-los, usando operações matemáticas ou métodos de preenchimento de dados faltantes para proporcionar uma testagem mais aprofundada e ampla das características do dados e de quais métodos produzem melhores resultados.

Para o problema de predição de crises financeiras, propõe-se fragmentar os períodos de amostragem em três grandes grupos definidos: (1) as instâncias onde se encontram o ano real de ocorrência das crises; (2) os quatro anos pós-crise; e finalmente (3), os anos que antecedem uma crise financeira ou a próxima crise financeira. Para (1) e (2) pode-se seguir a literatura estudada onde tais períodos podem ser excluídos da análise para evitar viés de crise BUSSIERE; FRATZSCHER (2006), uma vez que a economia está se recuperando da crise e, consequentemente, seus indicadores de crise (atributos do *dataset*) podem apresentar variações de recuperação (ajuste de preços, desempregos, inflação entre outros índices). Ainda, é interessante remover instâncias de períodos ruidosos, como as duas grandes guerras mundiais do século 20 e a grande depressão americana de 1929. Finalmente, serão realizadas operações nos atributos aos quais criam indicadores econômicos (descritos no capítulo 4) e a variável resposta indicadora de crise é ajustada para anos pré crises financeiras, uma vez que pretende-se prever uma crise anos antes.

Inicialmente, pode-se ajustar o *dataset* com apenas as instâncias que estão completas. Para este *dataset*, nomeamos de *bank*. No entanto, para aumentar os horizontes de conhecimento, pode-se utilizar técnicas para preencher variáveis faltantes, utilizando a classe da bibli-

oteca *sklearn*: *SimpleImputer*. Essa classe permite ampliar o número de dados. O resultado são 3 novos *datasets* definidos por cada de cada estratégia selecionada de preenchimento: *mean*, *median* e *most_frequent*, que respectivamente, preenchem cada conjunto a partir da média do atributo de cada país, da mediana e o valor mais frequente. Com os 4 *datasets* criados, pode-se partir para o tratamento de dados, ajustando cada conjunto de dados ao modelo para resultar em um desempenho comparativo superior quando avaliadas variáveis individualmente. A figura 15 representa o fluxo desta etapa de preparação de dados.

Figura 15 – Exemplo de classficação por máquinas de suporte vetorial



O Autor (2021)

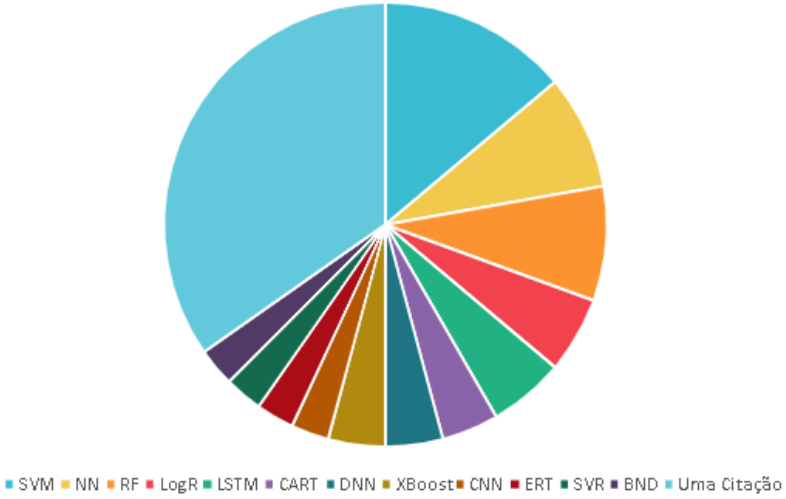
3.1.4 Etapa 4 - Aplicação dos Modelos Preditivos

Inicialmente, deve-se escolher os modelos preditivos para realizar avaliação preditiva. A partir dai, inclui-se no processo selecionar métodos que pesquisem padrões nos dados e decidam quais modelos e parâmetros podem ser mais apropriados, a exemplo da escolha de modelos com dados categóricos melhores do que modelos em vetores sobre reais. Dessa forma, na literatura estudada, destacam-se alguns métodos com maior incidência de citação e, por isso, são selecionados para a continuidade do estudo. São eles: SVM (10 vezes), NN e RF (6), LogR e LSTM (4).

Outro motivo pelo qual segue-se a lógica proposta para a seleção desses modelos preditivos é que uma vez que estes modelos são os mais citados na revisão sistemática, entende-se que possivelmente esses modelos evitam problemas de suavização como o *overffiting*. Salienta-se também que a maior parte dos modelos observados na revisão sistemática são citados apenas uma vez. No entanto, apesar da maior parte dos modelos serem descartados, os modelos selecionados representam 42% das citações revisadas. A figura 16 auxilia na visualização dos modelos preditivos para o problema de previsão de crises selecionados. Lê-se em sentido horário o mo-

delo SVM (azul) seguido dos modelos NN, RF, LogR e LSTM, até finalmente o conjunto de trabalhos citados uma única vez (maior parcela).

Figura 16 – Quantidade de citações dos modelos preditivos na literatura pesquisada



Fonte: O Autor (2021)

3.1.5 Etapa 5 - Comparação dos Modelos Selecionados

Nesta fase, objetiva-se interpretar os resultados e utilizar os padrões descobertos para incorporar esse conhecimento ao sistema de desempenho, tomando ações com base no conhecimento. Ainda, busca-se documentar, bem como verificar e resolver possíveis conflitos com o resultado previamente acreditado (ou extraído) do conhecimento. Para auxiliar a verificação são utilizadas métricas de avaliação, que são definidas e apresentadas anteriormente no capítulo 2.

3.1.6 Etapa 6 - Melhor Modelo

A partir da comparação dos resultados obtidos com as métricas avaliadas, busca-se entender os indicadores junto ao método que encontraram o melhor resultado. Usando os modelos e técnicas desenvolvidas, pode-se examinar as principais descobertas em maiores detalhes, especialmente, na interpretação computacional dos resultados nas métricas estabelecidas. Futuramente, o melhor resultado pode auxiliar outros estudos a identificar o risco de ocorrência de crises financeiras com antecedência e remediar de acordo com os sinais apresentados. Assim, pode-se fornecer mecanismos para identificação dos principais indicadores econômicos de previsões gerado por esses modelos, permitindo que os modelos de aprendizado de máquina sejam integrados a uma estrutura de tomada de decisão. Estas considerações estão detalhadas no próximo capítulo.

3.2 PLATAFORMAS DE *SOFTWARE* PARA *MACHINE* E *DEEP LEARNING*

Para que seja possível a coleta e manipulação dos dados selecionados são necessárias implementações de algoritmos a partir de bibliotecas, classes e plataformas de *software*. Tais algoritmos implementam a etapa de preparação dos dados, os modelos preditivos selecionados e as métricas de avaliação de desempenho que indicam o melhor modelo. Neste trabalho são propostos *scripts* utilizando linguagem de programação *Python* a partir de ferramentas como *Google Colab*, *Pandas*, *Keras*, *Scikit-lLearn* e entre outros. Para que se possa utilizar esses recursos e ferramentas basta criar um cadastro *online* no *gmail* que fornece uma autorização de uso da plataforma *google drive*. Na Figura 17) pode-se observar a integração de plataformas e bibliotecas divididas em quatro grandes áreas: (1) dados, manipulação de dados e tratamento dos atributos; (2) modelagem numérica e processamento de dados; (3) visualização dos resultados; e finalmente (4) o tema central deste trabalho utilizando a plataforma de implementação *Google Colab* que pode ser programada na linguagem de programação *Python*. Todos esses grupos são descritos a seguir, partindo da linguagem de programação selecionada, plataforma de programação, dados, manipulação de dados e visualização dos dados.

Figura 17 – Integração de Bibliotecas e Ferramentas

Fonte: O Autor (2021)

3.2.1 Linguagem *Python*

Uma das linguagens mais utilizadas em ciência de dados e enfim para aplicação em modelos preditivos é o *Python* (logo na Figura 18). Segundo COSTA (2020), a linguagem de programação ficou em primeiro no *ranking* de popularidade PYPL (*PopularitY of Programming Language*[2]). A implementação da linguagem *Python* iniciou em 1989 na Holanda e possui como uma das suas principais características não ser complexa, facilitando a ambientação, compreensão e aprendizado para diferentes aplicações.

Figura 18 – Logo da Linguagem *Python*

Fonte: PYTHON (2020)

O *Python* é uma linguagem de código livre, que pode ser aplicada em diferentes plataformas. Utilizando um arranjo de poucos caracteres especiais, possui uma identificação clara para marcação de bloco de código (identação), facilitando a sua leitura e possíveis manutenções que tenham de ser realizadas ao longo da construção código. Outra característica é que possui uma biblioteca padrão, contendo métodos e funções básicas que atendem desde o acesso a bancos de dados até a interface gráficas de iteração com o usuário. Além disso, suporta múltiplos paradigmas de programação como orientado a objetos, leva menos tempo de execução de códigos do que em outras linguagens e possui uma ampla disponibilidade de pacotes e bibliotecas, que a auxiliam na construção de algoritmos de aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*). Devido essas características, o *Python* acabou reunindo uma grande (e crescente) comunidade colaborativa de especialistas ao seu redor, MATOS (2020).

Conforme são realizados novos desenvolvimentos nos ambientes que suportam *Python*, a comunidade tende a colaborar nas discussões de problemas para auxiliar nas soluções de trabalhos apresentados por outros membros. Com isso posto, fica compreensível entender porque o *Python* é uma das linguagens mais usadas pra aplicação de códigos que contemplam métodos preditivos: é possível utilizar bibliotecas com módulos feitos em outras linguagens em um ambiente mais amigável conciliando performance e praticidade, facilitando a entrada de pesquisadores que muitas vezes não possuem computação como base de formação. Ainda, o pesquisador pode vir a adotar a linguagem *Python* por esta permitir que ele apresente os resultados de seus

[2] *PopularitY of Programming Language*: O índice de popularidade de linguagem de programação PYPL é criado analisando a frequência com que os tutoriais de linguagem são pesquisados na plataforma de pesquisa *Google*. Quanto mais um tutorial da linguagem selecionada é pesquisada, mais popular a linguagem é considerada. Os dados considerados são obtidos a partir do *Google Trends*. Para acessar o *ranking*: https://pypl.github.io/PYPL.html.

estudos e aplicações de forma clara, através do uso de bibliotecas que facilitam a criação de interfaces de visualização e enfim a interpretação dos dados que estão sendo estudados.

O *Python* tem uma grande variedade disponível de plataformas, bibliotecas e classes voltadas para as áreas de estudo com manipulação dados, sendo eficiente para aplicações com métodos preditivos. Para este estudo, inicialmente, destacam-se *Google Colab* (plataforma de programação), *Tensor Flow*, Pandas, Keras, *Scikit-learn*, *NumPy*, *SciPy*, *Seaborn* e *Pytorch* (breve apresentação nos capítulos subsequentes).

Por fim, a linguagem *Python* pode facilitar o trabalho de pesquisadores a obter resultados e descobertas a partir da programação e utilização dos algoritmos, (ACADEMY, 2020). No entanto, é importante ressaltar que a escolha de qual técnica deve ser utilizada para realizar a análise e o aprendizado de máquina pode depender do problema estudado em que se está tentando resolver.

3.2.2 Plataforma *Google Colab*

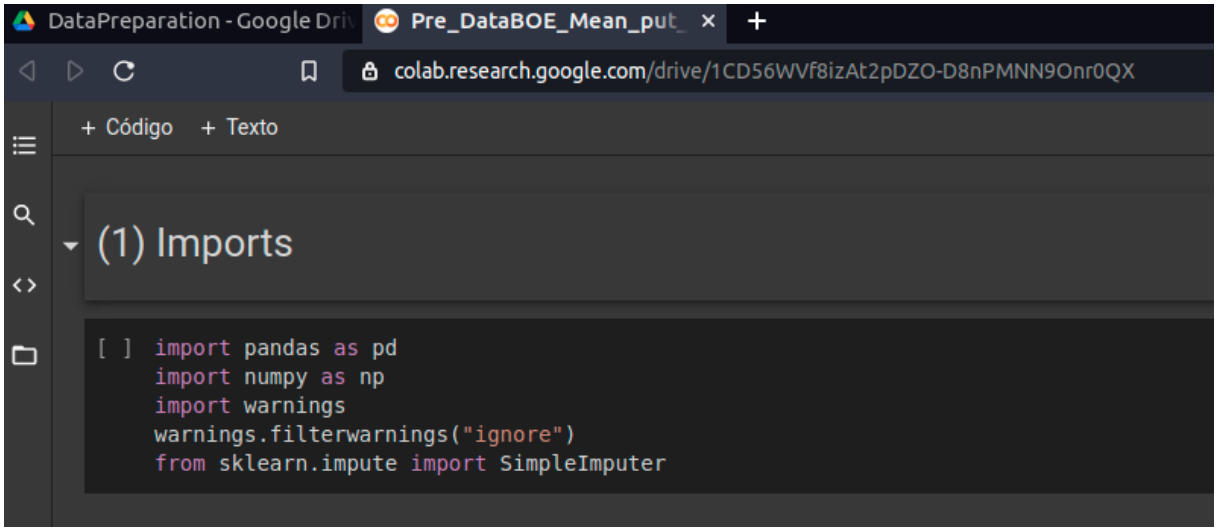
O *Google Colaboratory*[3], mais conhecido com *Google Colab*, é uma plataforma gratuita de código aberto que permite que usuários escrevam códigos na linguagem *Python* e os execute. Conforme o site oficial, os projetos são armazenados no *Google Drive* que permite os desenvolvedores a criar e compartilhar documentos com códigos, lógicas, equações visualizações e texto com outros usuários. Também, seus códigos podem ser carregados na plataforma *GitHub*[4]. Suas aplicações incluem refinamento e transformação de dados, simulações, modelagens estatísticas, *machine* e *deep learning*, entre outros a partir da integração de várias bibliotecas e ferramentas. Na Figura 19 pode-se observar um trecho de código de importação de bibliotecas como *Pandas* e *Numpy* que serão apresentados nas próximas subseções.

3.2.3 Bibliotecas de manipulação de dados e tratamento de atributos

Para realizar a etapa de pré processamento de dados, as bibliotecas *Pandas*[5] e *Pytorch*[6] são as ferramentas ideais. O *Pandas* (utilizado na implementação do presente trabalho) é uma biblioteca criada para a linguagem *Python* que oferece classes e estruturas de dados para análise preditiva e manipulação de dados. Seu uso contempla desde a visualização de todas instâncias [7], leitura e importação do *dataset*[8] . Destaca-se que é possível utilizar a biblioteca para manipular tabelas com grande quantidade de números e realizar operações em séries temporais. Internamente, possui métodos estatísticos que permitem agrupar, filtrar e combinar os dados. Para a

[3] Endereço de acesso: https://colab.research.google.com/notebooks/intro.ipynb#recent=true
[4] Endereço de acesso: https://github.com
[5] https://pandas.pydata.org/
[6] https://pytorch.org/
[7] Método: pd.set_option
[8] Método: pd.read_excel

Figura 19 – Importação de bibliotecas utilizando a plataforma *Google Colab*



Fonte: O Autor (2021)

análise de dados financeiros, a biblioteca pode auxiliar nos estudos uma vez que a base de dados inicialmente selecionada varia seus atributos ao longo do tempo.

As ferramentas que envolvem o ecossistema da *Pytorch* são voltadas principalmente para os campos de visão computacional e processamentos de NN para aprendizado de reforço. Além de utilizar a linguagem *Python*, pode ser utilizada para desenvolvimento na linguagem *C++*. Ainda, destaca-se que possui suporte em plataformas de nuvem, possuindo API completa relacionada para tratamento de dados para o modelo preditivo NN.

3.2.4 Bibliotecas de modelagem numérica e processamento de dados

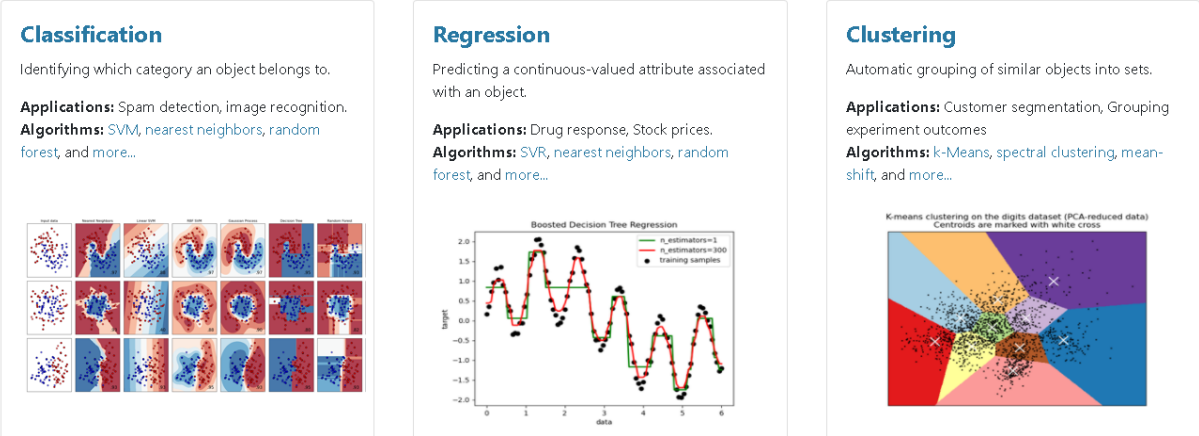
A etapa de implementação dos modelos selecionados demanda técnicas de modelagem numérica, tratamento estatístico e processamento de dados. Essas técnicas podem ser utilizadas a partir de classes importadas das bibliotecas *Scikit-learn*[9], *Numpy*[10], *Tensor Flow*[11], Keras[12] e *SciPy*[13]. Tais bibliotecas possuem diferentes particularidades que juntas podem trazer variadas comparações de modelos preditivos.

Partindo da biblioteca de código aberto *Scikit-learn* para a linguagem de programação *python*, disponibiliza algoritmos para tarefas dentre elas, classificações, regressão, *clustering*, redução de dimensionalidade, seleção de modelos e pré-processamento de dados e a partir de métodos de aprendizado de máquina. Pode-se visualizar algumas dessas aplicações na Figura 20. Inclui-se nesta biblioteca algoritmos relevantes para este trabalho de conclusão como SVM,

[9] https://scikit-learn.org/stable/
[10] https://numpy.org/
[11] https://www.tensorflow.org/
[12] https://keras.io/
[13] https://www.scipy.org/

RF, *K-means*. Ainda, destaca-se que ela é projetada para interagir com outras bibliotecas também desenvolvidas em *Python* para soluções numéricas e científicas, conforme o site oficial.

Figura 20 – Exemplos de aplicação da biblioteca *Scikit-learn*



Fonte: O Autor (2021)

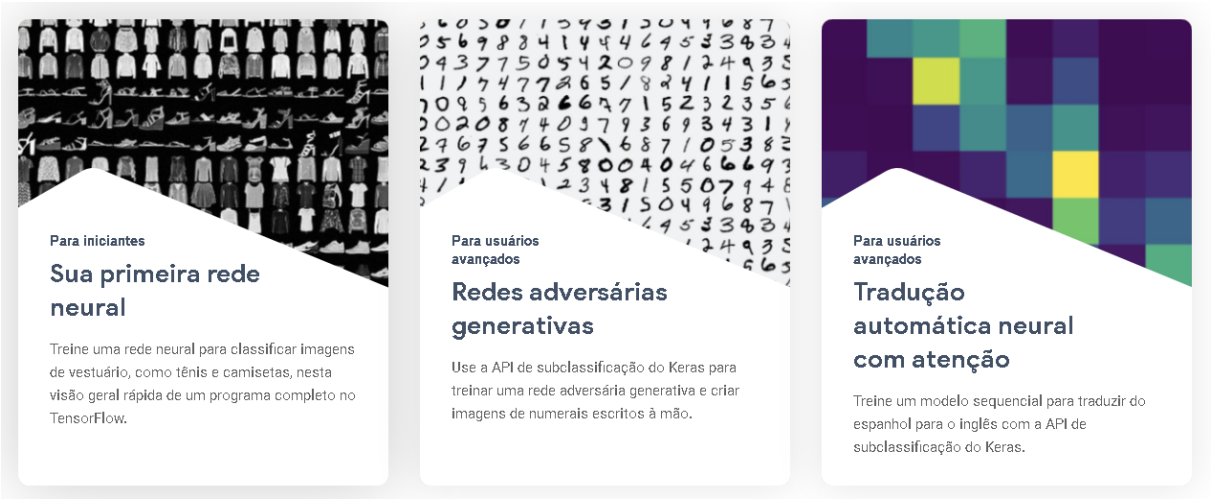
O *Numpy*, é outra biblioteca que possui classes para a aplicação na ferramenta *Google Colab* em *Python*. Geralmente, suas funcionalidades são comparadas ao *matlab*, uma vez que ambos são interpretadas e permitem ao usuário escrever programas simples, usando listas e matrizes com conjuntos de dados. Com o *Numpy*, pode-se construir algoritmos de *machine learning*, realizar operações de processamento de imagem, computação gráfica entre outras tarefas matemáticas.

O *Tensor Flow*, é uma iniciativa *Open Source*, onde tem suas aplicações destinadas para tarefas de *machine learning*. Desenvolvido pela gigante americana *Google*, utiliza grafos que podem ser usados para representar os nós dos modelos de NN, bem como as arestas que representam os tensores. O *Tensor Flow* trata *frameworks* para *Deep Learning* que possui uma grande variedade de modelos. Algumas das soluções oferecidas podem ser observadas na Figura 21.

Dentre as bibliotecas observadas, o *Tensor Flow* contém ferramentas para trabalhar com NN com várias partições de conjuntos de dados, o que e torna mais simples a transição de um protótipo de testes para um sistema de produção. Existe ainda a possibilidade de implantar e treinar modelos na nuvem, independente da linguagem de programação utilizada.

Com a atenção voltada para *deep learning*, a bilioteca *Keras*, tem suas aplicações desenvolvidas para trabalhar com NN que rodam sob a ferramenta *Tensor Flow*. Inicialmente seu objetivo é o de permitir experimentos rápidos com os algoritmos de aprendizado profundo (*deep learning*), ou seja, possibilita evoluir a pesquisa para o resultados com grande rapidez. A partir da linguagem *Python*, auxilia na implementação dos métodos preditivos que atendem as necessidades dos estudos. Além disso, a biblioteca contém as diretivas necessárias para realizar uma prototipagem rápida e fácil, suporte a CNN e LSTM.

Figura 21 – Soluções oferecidas pela biblioteca *Tensor Flow*



Fonte: O Autor (2021)

Por fim, o *SciPy* (Figura 22) apresenta recursos para soluções em problemas que envolvem álgebra linear, cálculo integral, probabilidades, processamento de sinais e imagens. Suas matrizes são baseadas nas bibliotecas oferecidas pela *Numpy*, fazendo parte do grande arranjo de aplicações de técnicas de *machine learning* para a linguagem *python*.

Figura 22 – Soluções oferecidas pela biblioteca *SciPy*

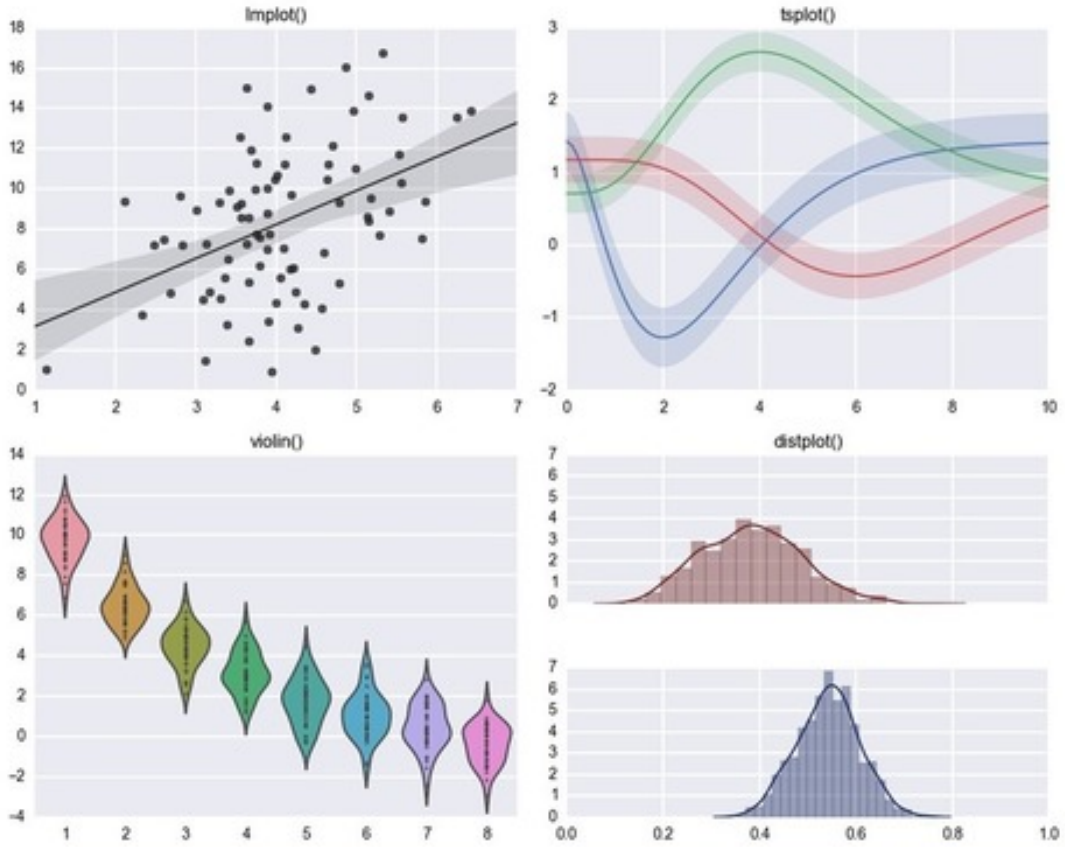
Fonte: FLAIR (2020)

3.2.5 Bibliotecas de visualização de resultados

Com o objetivo de obter uma melhor experiência e clareza na visualização de resultados, destacam-se duas bibliotecas que podem ser exploradas: *Seaborn*[14] e *Yellowbrick*[15].

A *Seaborn* se destaca, principalmente, por fornecer uma interface para desenhar gráficos estatísticos e informativos. Suas funções para construção gráficas operam em matrizes contendo conjuntos de dados e executam internamente um mapeamento para agregar as funções estatística necessárias para a produção dos mesmos. A Figura 23 apresenta diagramas de dispersão, enxames, diagramas de violino, entre outros que podem ser criados a partir da biblioteca.

Figura 23 – Técnicas de visualização disponíveis no *Seaborn*



Fonte: JOSHI (2018)

Por fim, a biblioteca *Yellowbrick* possui vários casos de uso, por exemplo: ajudar a avaliar a estabilidade e o valor preditivo dos modelos de aprendizado de máquina e melhorar a velocidade do fluxo de trabalho experimental, ferramentas visuais para monitorar o desempenho do modelo em aplicações com dados do mundo real, interpretação visual do comportamento do modelo no espaço de recursos de alta dimensão e ensinar/compreender uma grande variedade de algoritmos e métodos. A biblioteca *Yellowbrick* é uma plataforma de visualização de

[14] https://seaborn.pydata.org/
[15] https://www.scikit-yb.org/en/latest/

diagnóstico para aprendizado de máquina que permite que a condução o processo de seleção de melhor desempenho do modelo. Sua API baseada a partir de bibliotecas como *Scikit-learn*, possui o objetivo principal de auxiliar na visualização de resultados para interpretação. Os visualizadores permitem que os modelos vistos sejam ajustados e transformados como parte do processo, fornecendo diagnósticos visuais durante a transformação de dados. Pode-se destacar como métricas de visualização que podem ser construídas a partir da biblioteca, aplicações de curvas ROC/AUC e matrizes de confusão.

4 RESULTADOS DO ESTUDO DA PREDIÇÃO DE CRISES FINANCEIRAS

Os resultados são apresentados em 4 etapas. Primeiramente, descreve-se todo processo de preparação de dados conforme estudado, bem como as características dos atributos observados antes da preparação e pós preparação. Em segundo lugar, parte-se para os detalhes das métricas de avaliação aplicadas. Terceiro, são detalhados os resultados obtidos para cada método preditivo e *dataset* utilizado. Finalmente, são discutidos tais resultados para o modelo que apresentou os maiores valores nas métricas de avaliação de resultado: RF. Para encontrar os modelos construídos, desde a obtenção do *dataset* original, seguindo-se toda a metodologia deste trabalho, basta acessar o repositório na plataforma Github[1].

4.1 RESULTADOS DAS ETAPAS INICIAIS

Nesta seção apresenta-se os resultados obtidos nas etapa 1 (identificação de bases de dados), etapa 2 (seleção de atributos) e na etapa 3 (preparação dos dados). As etapas são importantes para a qualidade dos resultados obtidos. Nestas etapas foram aplicados diversos métodos e filtros a fim de identificar quais seriam as melhores alternativas. Neste sentido, observou-se que apenas um *dataset* poderia limitar os resultados obtidos. Por esta razão, decidiu-se implementar mais de um conjunto de dados, conforme descreve esta seção.

No primeiro momento foram analisadas as características do *dataset* original obtido a partir da publicação dos autores BLUWSTEIN *et al.* (2020). Tais características consistem em 29 atributos, com 2499 instâncias divididas para 17 países. A Tabela 8 descreve todos os atributos deste *dataset*, bem como indica a quantidade de instâncias com valores não nulos.

Foi observado que alguns atributos com valores faltantes. A análise realizada no *dataset*, em conjunto aos códigos do trabalho de referência, identificou a necessidade de uma preparação prévia dos dados, antes da aplicação dos métodos preditivos selecionados. Esta preparação de dados possui os estágios de eliminação de períodos históricos delicados (como as grandes guerras do século 20) e a eliminação de atributos sem relevância para o problema de predição de crises financeiras. Além disso, foi realizada a exclusão das instâncias no ano em que a crise ocorre. Isso porque, uma vez que se quer prever crises, foca-se nos dois anos anteriores da crise acontecer. Ou seja, a variável resposta de crise recebe o valor "1"nos dois anos anteriores a crise de fato ocorrer. Foi feita ainda a exclusão das 4 instâncias no ano pós crise (para evitar indicadores de crise que já ocorreram), conforme descrito em 3.1.2 e 3.1.3. Com estas etapas aprofundadas, percebeu-se que o atributo referenciado "*GDP*"(PIB em português) é de extrema importância, uma vez que ele é utilizado para realizar as operações em conjunto aos outros

[1] https://github.com/Sauthier1935/iafinancialcrisis

Tabela 8 – Características do *Dataset* original para aplicação dos algoritmos selecionados

Atributo	**Descrição**	**Nº Instâncias não nulas**
year	Ano dos dados	2499
country	País	2499
iso	identificador do país	2499
ifs	código do sistema financeiro	2499
pop	população	2499
rgdpmad	PIB real per capita (PPP)	2499
rgdppc	PIB real per capita (índice, 2005 = 100)	2499
rconpc	Consumo real per capita (índice, 2006 = 100)	2411
gdp	PIB	2474
iy	Proporção entre investimento/PIB	2279
cpi	Preços ao Consumidor	2249
ca	Conta corrente	2344
imports	Taxa te importação	2458
exports	Taxa de Exportação	2458
narrowm (nmon)	Dinheiro de facil liquidez	2435
money (bmon)	Todo o dinheiro que uma economia de um país possui	2349
stir (srate)	Taxa de juros de curto prazo	2351
ltrate (lrate)	Taxa de juros de longo prazo	2464
stocks (stock)	Preços das ações	2233
debtgdp (pdebt_gdp)	Proporção entre Dívida Pública/PIB	2322
revenue	Receitas do governo	2398
expenditure	Despesas do governo	2417
xrusd	Taxa de câmbio do dólar americano	2496
crisisJST (crisis)	Crises financeiras sistêmicas (0 ou 1)	2499
tloans (tloan)	Total de empréstimos ao setor privado	2311
tmort (mort)	Empréstimos hipotecários para o setor privado não financeiro	2186
thh (hloan)	Total de empréstimos às famílias	1309
tbus (bloan)	Total de empréstimos para empresas	1234
hpnom (hp)	Preços das casas	1906

Fonte: O Autor (2021)

atributos na criação do *dataset*.

Essa etapa de operações é inspirada no trabalho dos autores BLUWSTEIN *et al.* (2020), no qual são realizadas relações matemáticas econômicas: a inclinação da curva de juros (diferença de curto e taxas de juros de longo prazo); crédito (empréstimos ao setor privado não financeiro); preço de ações na bolsa; relação da dívida pública (crédito × taxa de juro de longo prazo sobre o PIB); consumo real per capita; são transformados em taxas percentuais dos índices dados. Para os demais atributos, ou seja, crédito, dinheiro, dívida pública, serviço da dívida, investimento e conta corrente, são diferenças em relação ao PIB. Para além destes atributos, são definidos dois

atributos globais, nomeados de crescimento do crédito global e a inclinação global da curva de rendimentos. Tais atributos são calculados para um par país-ano pelo crédito médio para o valor do PIB (inclinação média da curva de rendimento) em todos os países. Esses atributos calculados são amplamente explorados na literatura nas publicações dos autores (BORIO; LOWE, 2002; DREHMANN; BORIO; TSATSARONIS, 2011; SCHULARICK; TAYLOR, 2012; AIKMAN; HALDANE; NELSON, 2015). A Tabela 9 apresenta todos os atributos obtidos após o tratamento e suas respectivas descrições que serão utilizadas para aplicação dos modelos preditivos.

Tabela 9 – Descrição dos atributos

Atributo	**Descrição**
cpi_pdiff2	Percentual de preços ao consumidor
bmon_gdp_rdiff2	Diferença absoluta entre todo dinheiro da economia e PIB
stock_pdiff2	Percentual do preço de ações
cons_pdiff2	Percentual do consumo real per capta
pdebt_gdp_rdiff2	Proporção de dívida pública sobre o PIB
inv_gdp_rdiff2	Proporção de investimentos em relação ao PIB
ca_gdp_rdiff2	Soma total dos valores de contas corrente sobre o PIB
tloan_gdp_rdiff2	Total de empréstimos ao setor privado não financeiro sobre o PIB
tdbtserv_gdp_rdiff2	tloan multiplicado por lrate sobre o PIB
global_loan2	média percentual do Crescimento do crédito global excluindo o país selecionado em um horizonte de 2 anos
drate	Diferença entre stir e lrate
crisis	Variável categórica resposta de crise ou não crise
crisis_id	Todas as observações da que indicam a mesma crise recebem o mesmo ID Este ID é usado para o método de cross-validation para se certificar de que a mesma crise não está no conjunto de treinamento e teste.
year	Ano
iso	Identificador do País

Fonte: O Autor (2021)

Observando-se que o número de instâncias que possuíam atributos com valores nulos era significativo, optou-se por utilizar métodos para preenchimento dos dados faltantes. Assim, criou-se outros três *datasets* utilizando métodos diferentes para este processo de preenchimento. Para isso, foi utilizada a classe *sklearn.impute* que se encarrega de preencher os dados faltantes dos atributos indicados. Neste caso, vários atributos foram preenchidos, tais como: "rconpc", "gdp", "iy", "ca", "imports", "exports", "nmon"e "bmon". Estes e outros atributos, foram preenchidos considerando três parâmetros para o preenchimento (gerando 3 datasets distintos): *mean*: referente a média do atributo de cada país; *median*: referente a mediana do atributo de cada país; e, finalmente, *most_frequent*: referente ao valor mais frequente do atributo de cada país. O resultado da preparação de dados totalizou 4 *datasets* (original denominado *bank* e os três criados). A Tabela 10 descreve a identificação do *dataset* e as quantidades de atributos e instâncias de cada um. Note que o *dataset bank* contém menos instâncias pois aquelas que continham algum atributo nulo tiveram de ser deletadas.

Tabela 10 – Detalhamento dos *datasets*

Identificação	Nº de Atributos	Nº de Instâncias
bank	15	1249
mean	15	1714
median	15	1714
most_frequent	15	1714

Fonte: O Autor (2021)

Prontamente, após a criação dos *datasets* identificou-se ainda que dois atributos ("year"e "iso") não representavam características importantes para permanecerem no *dataset*. Portanto, os *datasets* finais ficaram com 13 atributos. Em seguida, observou-se que os *dataset* estavam desbalanceados. Uma vez que o problema de crises financeiras não é muito frequente, suas observações no período histórico do *dataset* original selecionado são limitadas. Na realidade, o atributo resposta indicador de crise (no *dataset* "crisis") estava dividido em 1154 ocorrências para não-crise (0) e 95 ocorrências para crise (1) no *dataset bank*; e para os demais *datasets* criados, 1491 ocorrências para não-crise e 129 ocorrências para crise. Para resolver este problema, a classe *imblearn.over_sampling* com o método *SMOTE* foi utilizada. O resultado obtido foi um aumento no número de instâncias para os *datasets* agora balanceados como pode ser visto na Tabela 11. Nela, apresenta-se o número de atributos e instâncias obtidas. Ainda, uma vez que as novas instâncias são adicionadas ao final do *dataset*, detectou-se uma sequência grande de instâncias com o atributo resposta (indicador de crises) com o valor 1. Sabendo-se que a ordem das instâncias no *dataset* podem afetar o processo de formação dos conjuntos de treino e teste, foi utilizado o método *shuffle* que randomiza as instâncias. Este método pode ser obtido na biblioteca *sklearn.utils*.

Tabela 11 – Descrição do *dataset* após aplicação do Método SMOTE

Nome Método	Nº de Atributos	Nº de Instâncias
bank	13	2308
mean	13	3148
median	13	3148
most_frequent	13	3148

Fonte: O Autor (2021)

Finalmente, com os *datasets* ajustados, partiu-se para a próxima etapa do trabalho.

4.2 RESULTADOS DA ETAPA 4 - APLICAÇÃO DE ALGORITMOS

Na etapa 4 foram definidos os algoritmos a serem testados, bem como os seus parâmetros. Como método de amostragem foi utilizada a validação cruzada (*cross-validation*). A validação cruzada contempla em ajustar os modelos selecionados para os dados do *dataset* em seleções aleatórias, chamados de conjuntos de treinamento e teste. Para o presente trabalho

particionou-se o *dataset* usando o parâmetro $cv = 10$ no método de *cross-validation*. Além disso, para realizar testes finais foram utilizadas técnicas próprias da área baseadas em hiper-parâmetros como identificar as melhores configurações para cada método.

Pode-se definir hiper-parâmetros como os valores que controlam o próprio processo de treinamento (DEVELOPERS, 2021). Em outras palavras, os hiper-parâmetros estabelecem quais faixas de valores de referência são utilizadas para treinar o modelo preditivo selecionado. O ajuste desses valores é realizado a partir de vários testes denominados episódios de treinamento. Cada episódio é uma execução completa do modelo preditivo com o *dataset* selecionado com as parametrizações definidas dentro dos limites especificados. O método avaliativo retorna valores para cada execução (com parametrização dentro das faixas estabelecidas), o que permite visualizar a melhor configuração para cada método. No presente trabalho, foram executados diversos métodos a fim de obter-se os melhores resultados de avaliação possíveis. Para definir os melhores parâmetros de cada método foram utilizadas classes da biblioteca *sklearn*: *GridSearchCV*, *train_test_split* (parâmetro *test_size* definido para o valor 0.75) e *classification_report*.

Os hiper-parâmetros testados para LogR são os seguintes: para "*penalty*", elasticnet, L1, L2, none; "*tol*", a faixa de valores entre 1 e 0.000001; para "*solver*", newton-cg, lbfgs, sag, saga; para "*multi_class*", auto, ovr e multinomial; "*C*", os valores 0.001, 0.01, 0.1 e 1; e finalmente como "*max-iter*", os valores 50,100,300,500,1000; onde melhores parâmetros observados para cada *dataset* estão apresentados na Tabela 12, sendo que observa-se que o parâmetro *solver* é o mesmo para todos os *datasets*.

Tabela 12 – Parâmetros definidos para LogR

Dataset	C	max_iter	multi_class	penalty	solver	tol
bank	0.001	50	multinomial	none	newton-cg	0.001
mean	0.01	300	auto	l2	newton-cg	1
median	0.001	100	auto	l2	newton-cg	1
most_frequent	0.001	50	auto	none	newton-cg	1

Fonte: O Autor (2021)

Para a definição dos hiper-parâmetros de SVM, foram testados os seguintes: (1) "*kernel*", rbf, sigmoid, poly; (2) "*gamma*", a faixa de valores entre 1 e 0.0000001; (3) "*decision_function_shape*", ovo e ovr; (4) "*C*, os valores 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500 e 5000; (5) "*tol*", a faixa de valores entre 1 e 0.000001; em que os melhores parâmetros para cada *dataset* estão apresentados na Tabela 13. Observou-se que a maior parte dos parâmetros não variaram nos *dataset*.

Foram testadas várias combinações de neurônios para NN, em conjuntos de 5 e 10 neurônios, no intervalo de 10 até 100. O melhor resultado para o parâmetro "*hidden_layer_sizes*"(*hly*), foi o conjunto de 5 camadas de 50 neurônios cada (5x50). Com este parâmetro definido, foi utilizado o método *GridSearchCV* para encontrar os demais melhores parâmetros do método pre-

Tabela 13 – Parâmetros definidos para SVM

Dataset	Kernel	gamma	decision_function_shape	C	tol
bank	rbf	0.001	ovo	5000	1
mean	rbf	0.001	ovo	2500	1
median	rbf	0.001	ovo	5000	1
most_frequent	rbf	0.001	ovo	5000	1

Fonte: O Autor (2021)

ditivo. São eles: "*activation*", identity, logistic (obteve o melhor resultado em todos os *datasets*), tanh e relu; "*alpha*", a faixa de valores entre 0.01 e 0.0000001; "*learning_rate*", constant, invscaling, adaptive; "*momentum*", os valores 0.5, 0.85, 0.9, 0.95 e 0.99; e finalmente "*max_iter*", valores entre 500 e 5000. Os melhores resultados estão na Tabela 14, a seguir.

Tabela 14 – Parâmetros definidos para NN

Dataset	hly	activation	alpha	learning_rate	momentum	max_iter
bank	5x50	logistic	0.0000001	adaptive	0.95	2500
mean	5x50	logistic	0.0000001	constant	0.90	2500
median	5x50	logistic	0.000001	adaptive	0.95	3000
most_frequent	5x50	logistic	0.0001	adaptive	0.99	3000

Fonte: O Autor (2021)

Para RF os arranjos de hiper-parâmetros testados foram "*Criterion*", gini e entropy; "*max_depth*", valores 5, 10,13, 15, 17, 20, 23, 25 e "*none*"; "*max_features*", auto, sqrt e log2; "*min_samples_leaf*"abreviado para "*msl*", os valores 1,2,3,4,5,6,7,8; "*min_samples_split*"abreviado para "*msp*", os valores 2, 5 e 10; e finalmente "*n_estimators*", os valores 100, 200, 300, 500, 750,1000,1250,1300, 1400, 1450, 1500 1525, 1550, 1575, 1600 ,1700, 1800 e 2000. Assim, a próxima Tabela 15, todas as combinações por *dataset* obtidas. Observa-se que os parâmetros *Criterion* e *min_samples_leaf* não se alteraram na obtenção dos melhores resultado.

Tabela 15 – Parâmetros definidos para RF

Dataset	Criterion	max_depth	max_features	msl	msp	n_estimators
bank	Gini	none	sqrt	1	2	1000
mean	Gini	25	sqrt	1	2	1525
median	Gini	none	sqrt	1	2	1575
most_frequent	Gini	23	auto	1	2	1000

Fonte: O Autor (2021)

Finalmente, para o método LSTM foram testados 5 parâmetros. Além disso, por se tratar de um método de *deep learning*, apresenta-se os código do modelo ajustado que foi construído, bem como a descrição de cada camada do LSTM. A Figura 24 apresenta a construção do modelo LSTM. O modelo obtido partiu de diversas pesquisas realizadas na *internet* que após

vários testes, foi validado. O modelo original pode ser encontrado na plataforma *StackOverflow*[2]. No entanto, para o problema do presente trabalho, foram necessárias algumas alterações. Essas alterações foram fundamentadas a partir da documentação biblioteca de código aberto *TensorFlow*[3], onde foram testados os parâmetros na criação do modelo: *Hidden_nodes*, os valores 5, 10, 15; *optimizer*, sgd, adam e rmsprop (observado que que o otimizador adam obteve superioridade em relação aos outros); e *loss*, que foi ajustado para *binary_crossentropy*, uma vez que este parâmero é utilizado para classificações que envolvem escolhas binárias (sim ou não; A ou B; 0 ou 1).

Figura 24 – Modelo LSTM construído

```
[ ]  def create_model():
       model = Sequential()
       model.add(LSTM( Hidden_nodes ,return_sequences=True))
       model.add(Activation('relu'))
       model.add(Dropout(0.2))
       model.add(LSTM(3))
       model.add(Activation('relu'))
       model.add(Dense(1))
       model.add(Activation('sigmoid'))
       model.compile(optimizer= Optimizer,loss= Loss ,metrics=['accuracy'])
       model.build(dfx_median.shape)
       print(model.summary())
       return model

lstm_scikitlearn =  KerasClassifier(build_fn=create_model, epochs= Epochs,
                                    batch_size = Batch_size, verbose=0)
```

O Autor (2021)

Assim, para LSTM, após pesquisas realizadas e o modelo ajustado, outros dois parâmetros também foram testados, *epochs* na faixa de valores entre 10 a 65; e finalmente *batch_size*, nos valores de 10 até 100. A seguir, na Tabela 16 pode-se observar os melhores valores de cada parâmetro obtido.

Na próxima seção são apresentados os resultados obtidos por cada método ajustado pelo seus melhores parâmetros.

4.3 RESULTADOS DA ETAPA 5 - COMPARAÇÃO DOS MODELOS

Nesta etapa, foi obtido o valor das métricas de desempenho preditivo dos diferentes modelos selecionados. Todos os modelos visam prever a ocorrência futura de uma crise financeira.

[2] Endereço que o modelo foi obtido: https://stackoverflow.com/questions/50840001/binary-classification-in-keras-and-lstm.

[3] A documentação pode ser obtida no seguinte endereço: https://www.tensorflow.org/guide/keras/rnn

Tabela 16 – Parâmetros definidos para LSTM

Dataset	Hidden_nodes	optimizer	loss	Epochs	batch_size
bank	15	adam	binary_crossentropy	50	20
mean	10	adam	binary_crossentropy	55	20
median	15	adam	binary_crossentropy	55	13
most_frequent	10	adam	binary_crossentropy	50	10

Fonte: O Autor (2021)

No entanto, a taxa de acerto é medida a partir de instâncias que foram previstas corretamente tanto em períodos de não crise (variável resposta 0) e crise (variável resposta 1). Pode-se avaliar o desempenho de cada modelo a partir das métricas acurácia, precisão, *recall* e *F1-Score*. O modelo perfeito obteria a pontuação 100 para a acurácia e pontuação 1 nas demais métricas citadas. As tabelas dessa seção mostram o desempenho de todos todos os modelos tiveram a partir dos *datasets* selecionados. Inicialmente, pode se observar que os melhores resultados individuais de cada método preditivo foram obtidos a partir dos *datasets mean* (três ocorrências) e *bank* (duas ocorrências). Ainda, todos os resultados mostrados são baseados em validação cruzada (*cross-validation*), seguindo a literatura estudada BLUWSTEIN *et al.* (2020) que por sua vez utilizou o parâmetro $cv = 5$. A tabulação dos resultados para cada método preditivo parte do *dataset* e resultado da acurácia. Após, as colunas são respectivamente: precisão, *Recall* e *F1-Score*. Finalmente, esta seção termina com a útima métrica selecionada: a curva ROC. Por sua vez, a curva ROC obtida é comparada aos resultados obtidos com a literatura. Todavia, o método utilizado de separação de dados de treino e teste foi o *train_test_split*, obtido através biblioteca *sklearn.model_selection*.

A Tabela 17 mostra os melhores resultados obtidos para para LogR, respectivamente, para cada *dataset*. Observou-se que o método preditivo LogR teve o desempenho baixo em nos três *datasets* com preenchimento de dados (*mean*,*median* e *most_frequent*). Ainda, percebeu-se uma grande queda no desempenho quando comparamos os resultados do método entre os *datasets*. Os resultados de *mean* e *most_frequent* obtiveram pouco mais de 50 pontos para acurácia.

Tabela 17 – Métricas de resultados obtidos para LogR

Dataset	Acurácia	Precisão	Recall	F1-Score
bank	81.195840	0.806943	0.825823	0.816274
mean	52.731893	0.850000	0.057009	0.106851
median	65.406607	0.722879	0.451375	0.555739
most_frequent	50.158813	0.857143	0.008048	0.015947

Fonte: O Autor (2021)

O método preditivo SVM apresentou valores de acurácia próximos (entre 85 e 88 pontos). Concentrando-se em avaliar as melhores métricas, os valores obtidos com o método ficaram em segundo lugar, como pode ser observado na Tabela 18.

Tabela 18 – Métricas de resultados obtidos para SVM

Dataset	Acurácia	Precisão	Recall	F1-Score
bank	87.348353	0.861411	0.899480	0.880034
mean	85.228716	0.881139	0.830315	0.854972
median	86.149936	0.864865	0.879946	0.872340
most_frequent	85.641677	0.876740	0.887324	0.882000

Fonte: O Autor (2021)

Em termos de métricas de desempenho, observamos que na Tabela 19, o modelo preditivo NN obteve resultados muito próximo nas acurácias (variação de pouco maior que 3 pontos) e que o *dataset* que obteve o maior valor na métrica de acurácia foi o *dataset* com preenchimento de células nulas com o parâmetro *mean*.

Tabela 19 – Métricas de resultados obtidos para NN

Dataset	Acurácia	Precisão	Recall	F1-Score
bank	76.559792	0.986555	0.508666	0.671241
mean	78.684879	0.841699	0.731053	0.782484
median	76.365946	0.949398	0.528504	0.679018
most_frequent	73.411689	0.838194	0.597586	0.697729

Fonte: O Autor (2021)

Dentre todos os modelos selecionados, a Tabela 20 mostra a superioridade nos resultados obtidos em todos os *datasets* nos modelos preditivos selecionados em RF. Além dos resultados na acurácia serem próximos (menos de 2 pontos), obtivemos uma superioridade nítida em todas as métricas de resultado, em comparação ao segundo colocado (SVM). Ainda, os *datasets mean* e *median* obtiveram maiores métricas nos valores na acurácia.

Tabela 20 – Métricas de resultados obtidos para RF

Dataset	Acurácia	Precisão	Recall	F1-Score
bank	96.100519	0.961905	0.962738	0.962321
mean	97.518443	0.980232	0.964453	0.972279
median	96.981891	0.972900	0.963112	0.967981
most_frequent	95.741113	0.965424	0.955064	0.960216

Fonte: O Autor (2021)

Por fim, o único modelo de *deep learning* selecionado, LSTM é apresentado na Tabela 21. A primeira impressão a ser destacada é que as métricas obtidas nos valores da acurácia com *datasets* com um número maior de instâncias obtiveram valores maiores. Isto quer dizer que para os testes realizados, os *datasets mean*, *median* e *most_frequent* obtiveram métricas com valores superiores ao *bank* (apesar de serem muito inferiores a outros modelos comparados).

Tabela 21 – Métricas de resultados obtidos para LSTM

Dataset	Acurácia	Precisão	Recall	F1-Score
bank	71.880415	0.978151	0.504333	0.665523
mean	73.910127	0.924855	0.536553	0.679117
median	73.809523	0.945806	0.491616	0.646955
most_frequent	73.071763	0.963614	0.515091	0.671329

Fonte: O Autor (2021)

Finalmente, como todos os modelos visam prever a ocorrência de uma crise financeira, pode-se avaliar seu desempenho utilizando a curva ROC (*Receiver Operating Characteristic*). A Figura 25 apresenta os gráficos das curvas ROC obtidas para os modelos selecionados para o *dataset bank*, onde cada modelo é indicado, respectivamente: *LogisticRegression* (LogR, em azul); *SVC* (SVM, em laranja); *MLPClassifier* (NN, em verde); *KerasClassifier* (LSTM, em vermelho); e por fim *RandomForestClassifier* (RF, em roxo). O eixo vertical do gráfico mostra a taxa de verdadeiro positivo, também conhecida como taxa de acerto, definida como a proporção de ocorrências positivas (crises) corretamente identificadas. Já eixo horizontal mostra a taxa de falsos positivos, ou seja, taxa de falsos alarmes indicadores de crises, definido como a proporção de ocorrências negativas (não crises) incorretamente identificadas como positivas (crise). Neste contexto, um modelo perfeito obteria uma taxa de acerto para 1 e uma taxa de falso alarme de 0. O desempenho geral de um modelo no espaço ROC pode ser resumido pela area sob a curva (AUC), conforme descrito na literatura estudada. Pode-se observar que o modelo RF obteve o maior resultado na curva ROC, indicando sua superioridade em comparação aos outros modelos preditivos estudados. É importante ressaltar que para o processo de construção da curva ROC foi utilizada classe *train_test_split* para obter vários pontos de teste na curva, uma vez que o *cross-validation* cria apenas um ponto no espaço para visualização. Tal entendimento é importante uma vez que alguns modelos preditivos obtiveram desempenho maior esperado visível na curva ROC, tal como o modelo LogR.

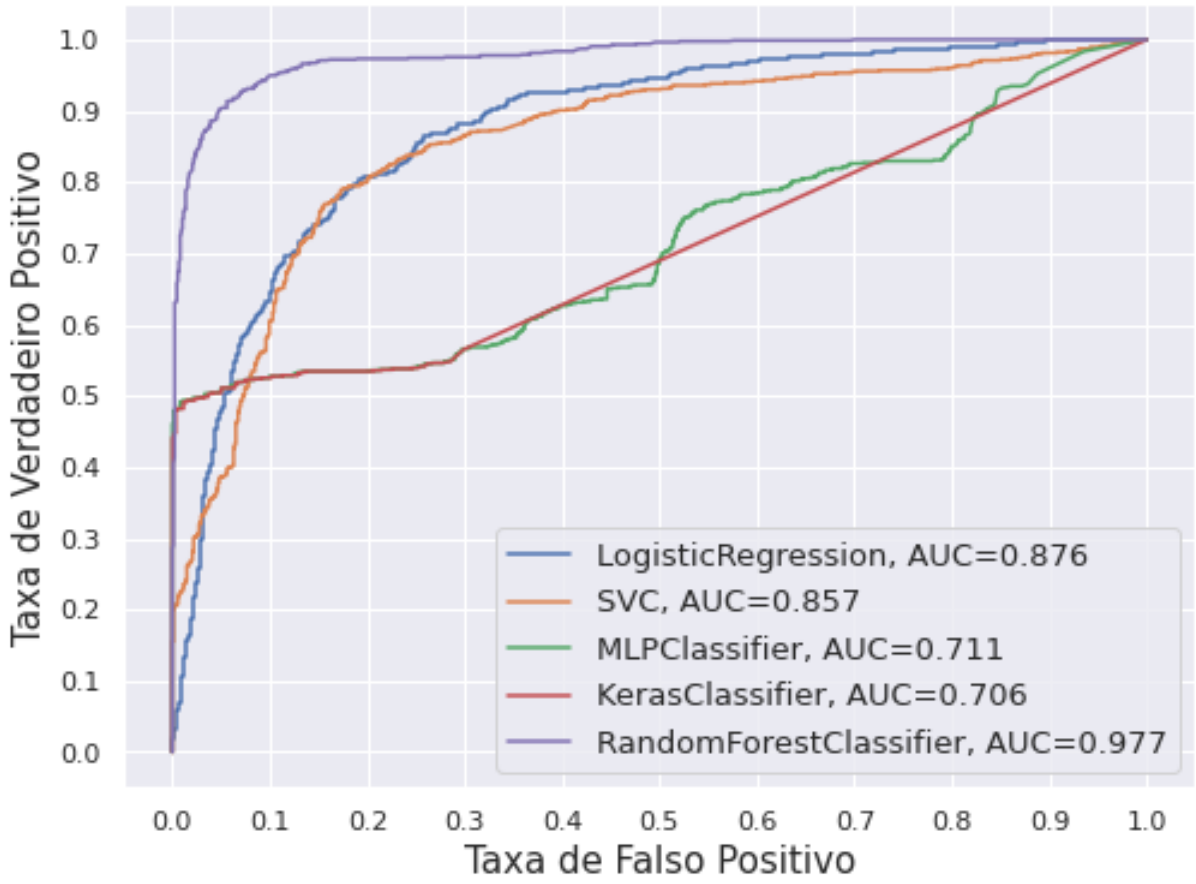
Assumindo o *dataset* original *bank*, pode-se comparar a área sobre a curva (AUC) com os trabalhos descritos na literatura que possuem melhores resultados. A Tabela 22) compara os resultados da curva AUC obtida no presente trabalho e a curva obtida para no trabalho BLUWSTEIN *et al.* (2020).

Tabela 22 – Comparação entre resultados obtidos com o *dataset Bank*

Método Preditivo	AUC - TCC	AUC - Trabalhos relacionados
LogR	0.876	0.822
SVM	0.857	0.832
NN	0.711	0.829
RF	0.977	0.855
LSTM	0.706	Sem resultado de comparação

Fonte: O Autor (2021)

Figura 25 – Curva ROC para os modelos preditivos selecionados



O Autor (2021)

4.4 RESULTADOS DA ETAPA 6 - MELHOR MODELO

Para melhorar a compreensão dos nossos resultados e dando continuidade as etapas definidas do trabalho, examinamos o modelo de melhor desempenho - RF (*Random Forests*). A principal métrica utilizada para apresentar os resultados foi a acurácia, que é uma das mais observadas na literatura estudada para finalidade de encontrar modelos que apresentam as melhores métricas. No entanto, para as outras métricas selecionadas, RF também obteve resultados superiores em comparação a outros métodos. Ainda, pode-se observar na tabela da seção anterior que os resultados foram consistentes independente do *dataset*, sendo que a variação da acurácia entre os *datasets* foram pouco maiores que 2 pontos percentuais. Pode-se perceber que os parâmetros *Criterion*, *min_samples_leaf* e *min_samples_split* estão bem definidos quanto a saída, dentro do arranjo de valores explorados para cada um desses parâmetros. As descobertas utilizando vários *datasets* sugerem que o número de instâncias dependendo a estratégia de preenchimento aumentaram o desempenho do método preditivo, principalmente utilizando as técnicas *mean* e *median*. Todavia, devido aos valores das métricas observadas serem próximos, não se pode definir qual *dataset* é o melhor para o problema definido. Pode-se perceber tam-

bém que não existe um padrão definido para obter a melhor acurácia utilizando os parâmetros *max_depth* e *max_features*. No entanto, para *n_estimators* pode-se observar que os melhores obtidos estão na na faixa de valores entre 1000 e 1600.

4.5 CONSIDERAÇÕES FINAIS

Chegando ao término deste trabalho, é importante tecer algumas considerações finais, tanto em relação ao processo quanto aos resultados. O processo de aprendizado de máquina compreende várias etapas, todas extremamente importantes para alcançar resultados satisfatórios em termos de aprendizado. Em primeiro lugar, observamos que a etapa de preparação e transformação dos dados é fundamental. Em conjunto à literatura estudada, está claro que para o cenário de predição de crises financeiras utilizando modelos preditivos, os dados brutos (*dataset* Macrohistória Jordà-Schularick-Taylor) não são suficientes para a construção de um modelo preditivo confiável. Por ser um problema complexo, são necessárias que novas variáveis sejam construídas, formuladas e combinadas à partir destes dados brutos. Ainda, como este presente trabalho tomou como base na etapa de preparação os estudos dos autores BLUWSTEIN *et al.* (2020), pode-se validar variáveis e dados de tais trabalhos precedentes. Em segundo lugar, em termos do processo que foi desenvolvido, nossos resultados indicam que as técnicas de *machine learning* operam bem sobre esses dados construindo modelos e representações coerentes com os trabalhos relacionados da literatura.

No intuito de buscar melhores resultados, comparativamente a literatura, foram criados neste trabalho outros três *datastes*, uma vez que entendemos que a diversidade de dados, expressões outras representações dos cenários de crise que fazem parte de um estudo exploratório. Ao longo deste estudo esses *datasets* desenvolvidos com técnicas de preenchimento se revelaram satisfatórios, uma vez que para alguns modelos preditivos, superaram os valores nas métricas em comparação ao *dataset bank*. Para NN, RF e LSTM, o *dataset* que obteve o melhor resultado entre o modelo, foi o com a técnica de preenchimento *mean*. Ainda, para o modelo preditivo RF não se obteve resultados melhores do que aqueles obtidos com o *dataset mean*, atingindo uma acurácia de 97.51 pontos percentuais. Também, pode-se observar que apesar dos resultados baixos para o modelo de *deep learning* LSTM, os *datasets* com maior quantidade de dados superaram o *dataset bank*. Em termos dos modelos preditivos selecionados, concluímos que o modelo RF é o mais apropriado para o problema de predição de crises financeiras, não envolvendo a necessidade de *deep learning* para obter melhores resultados. Sendo assim, visto a grande superioridade nas métricas obtidas, podemos concluir que o modelo LSTM, de acordo com os parâmetros selecionados, não se mostra necessário para problemas neste cenário.

Por fim, em terceiro lugar, destacamos que por se tratar de um trabalho exploratório, diversos parâmetros para buscar os melhores resultados nos modelos preditivos selecionado foram testados. Comparativamente, avançamos as buscas por melhores parâmetros em todos os modelos: por exemplo, enquanto RF nos trabalhos relacionados possui parâmetros fixados

nos modelos tais como *n_estimators* definidos com valor único, testamos uma faixa de valores para o parâmetro. Assim, esta etapa exploratória por parâmetros que apresentassem melhores resultados, apesar de demorada, nos levou a ter melhores escolhas (podendo filtrar) quanto a construção dos modelos preditivos.

5 CONCLUSÃO

Neste trabalho de pesquisa foram apresentados sistemas preditivos partindo-se de métodos de *machine learning* e *deep learning*. Como problema preditivo, considerou-se a predição de crises financeiras. Neste sentido, os modelos preditivos aplicados podem ser utilizados para identificar riscos de investimentos, avaliar resultados de decisões de equipes econômicas, tomadas de decisão e enfim tentar antecipar vários eventos que podem alterar a forma com a qual o sistema econômico funciona.

A partir do processo de revisão sistemática, pode-se identificar que já existem sistemas preditivos nessa área, onde os métodos mais utilizados são LogR, SVM, NN, RF e LSTM. No entanto, apesar dos resultados obtidos nos trabalhos relacionados, também observou-se que existe margem para melhorias. A partir daí, se identificou também que existem vários conjuntos de dados, mas um conjunto de dados de extrema importância é o *dataset bank*[1]. Destaca-se que uma etapa fundamental para se obter êxito na implementação dos modelos é a definição dos dados a serem analisados e a sua preparação, principalmente, quanto a etapa de seleção e criação de atributos que de fato são utilizados para o problema de predição de crises. Porém, o conjunto de dados bruto apresenta elementos que podem ser melhorados. No intuito de fazer essas verificações, foram criados três outros *datasets* utilizando técnicas de preenchimento: *mean*, *median* e *most_frequent*. Esses *datasets* revelaram-se importantes, uma vez que com o aumento do número de instâncias, alguns modelos obtiveram valores nas métricas de avaliação superiores ao *dataset bank*. Após a constituição e preprocessamento dos *datasets*, partiu-se para a aplicação de métodos preditivos.

Seguindo-se todas as etapas propostas, finalmente, chega-se na construção de modelos preditivos selecionados previamente (LogR, SVM, NN, RF e LSTM) que incluem a função de analisar o conjunto de dados selecionados. Para se extrair o melhor desempenho de cada modelo, técnicas de para seleção dos melhores parâmetros foram utilizadas. Esta etapa de seleção revelou-se de extrema importância, uma vez que pode-se explorar várias combinações de faixas de valores para esses parâmetros. Com esses parâmetros ajustados, finalmente, pode-se obter as métricas de avaliação cada modelo. Neste momento, constatou-se que a etapa exploração possibilitou obter melhores valores nas métricas quando comparado aos resultados da literatura. Comparativamente, após testes e avaliações, o método mais relevante para o presente trabalho é o RF.

Assim, a análise dos dados financeiros utilizando ferramentas de *machine learning*, em especial as *Randon Forests* (RF), é um ponto a ser exaltado, uma vez que a classificação pode trazer maior precisão nos conjuntos selecionados, independente do *dataset* explorado neste trabalho. Além disso, para os métodos utilizados para previsão de crises financeiras podem

[1] *Dataset* preprocessado obtido à partir do banco de dados de Macrohistória Jordà-Schularick-Taylor

possibilitar que as futuras crises sejam minimizadas e remediadas, permitindo que os agentes econômicos, futuramente, possam minimizar suas perdas e impactos diante a sociedade.

Por fim, no aspecto de previsão de crises financeiras, deve-se reforçar que este tema trata de um problema complexo, com várias variáveis dependentes da ação humana. Por isso, diante dos objetivos definidos neste trabalho de conclusão, conclui-se que foi possível atingir os pontos propostos. Crises financeiras podem influenciar nossa sociedade como um todo e é um tema de extrema importância para a saúde financeira de pessoas, governos e instituições prevê-las e remedia-las da melhor forma com antecedência.

5.1 CONTRIBUIÇÕES

A pesquisa realizada buscou atender as lacunas de diversos aspectos do ponto de vista dos modelos preditivos. Esses modelos preditivos estudados, selecionados e aplicados neste trabalho tinham como objetivo agregar previsões à pesquisa e também auxiliar na tomada de decisão para instituições financeiras, empreendimentos e governos que venham a ter a demanda para avaliação de riscos. A proposta de desenvolvimento por completo deste trouxe diversas dificuldades como a preparação dos dados (uma vez que crises financeiras são eventos raros ao longo da história, limitando a quantidade de variáveis resposta nesse sentido), que foram vencidas com a aplicação de conceitos e bibliotecas de programação adequadas, amplamente testadas durante o desenvolvimento. Outro desafio que pode-se ressaltar, foi a exploração de novos conjuntos de dados, criando-se novos *datasets* para ampliar a complexidade das previsões. Por fim, ao buscarmos os melhores parâmetros para os modelos preditivos selecionados, nos deparamos com infinitas possibilidades, o que pode servir de gancho para continuarmos pesquisas na área conforme a próxima seção.

A linguagem de programação escolhida para a implementação dos modelos, em conjunto a bibliotecas e ferramentas são, a primeira vista, de excepcional ajuda, uma vez que há um grande repositório de conteúdos na *internet* para elaboração de diferentes aspectos que impactam no resultado final de cada previsão. Quanto aos resultados obtidos até o momento se colocam como satisfatórios, visto que a utilização dos temas abordados - modelos preditivos, mercados, ciclos econômicos e enfim as crises que podem ocorrer na economia - aplicados às engenharias por meio de linguagem de programação apresentam um grande potencial para simplificar e facilitar análises futuras que buscam prever e remediar situações indesejáveis na sociedade como um todo.

Por fim, este trabalho também pode ser comparado a outras contribuições da área. O de maior semelhança, desenvolvido pelos pesquisados do banco central da Inglaterra (*Bank of England*), BLUWSTEIN *et al.* (2020). Neste projeto em questão, foram explorados outros modelos preditivos para o problema de predição de crises financeiras, tendo como o modelo que obteve as maiores métricas o modelo de árvores extremamente aleatórias (ERT). No entanto, devido

a busca aprofundada pelos melhores parâmetros, o modelo RF do presente trabalho obteve métricas de avaliação com valores superiores quando comparado ao modelo publicado. Enfim, os resultados em comparação a literatura são considerados similares, indicando que o modelo RF obteve o maior grau de confiabilidade e desempenho.

5.2 TRABALHOS FUTUROS

Para a continuação do trabalho desenvolvido, deve-se seguir a metodologia proposta no capítulo terceiro desta investigação para o problema de detecção de crises financeiras apresentado, a fim de aplicar os métodos preditivos e classificá-los conforme seu desempenho preditivo. Outras observações que poderão ser realizadas com base nos resultados obtidos são a avaliação da relevância de cada atributo no conjunto de dados utilizado; bem como aprimoramentos e refinamentos na etapa de testes de dados onde possa ser realizada uma comparação com outros classificadores para que estes sirvam como parâmetro. Assim, observa-se que, para uma posterior aplicação do trabalho, devem ser utilizados os métodos apresentados no capítulo dois, que são classificados a partir de métodos de aprendizado profundo (*deep learning*), avaliados através das classificações das variáveis resultadas de acurácia e área sob a curva na implementação deste trabalho. Outro ponto de que deve ser ressaltado é que novos parâmetros no universo de possibilidades de cada modelo devem ser testados uma vez que uma simples mudança destes pode impactar na melhoria na obtenção das métricas de resultado. No modelo LSTM, especificamente, pode se proposta novas construções de modelos diferentes do que o modelo utilizado.

Outros pontos que podem ser propostos para a continuidade do presente trabalho podem ser ligados a outros problemas que envolvem as áreas da economia e engenharia, que são listados à seguir.

- Avaliação de risco de empréstimos por parte do emprestador
- Avaliação de risco de empréstimos por parte do tomador de crédito
- Avaliação de risco na tomada de decisão de investimentos em ativos financeiros
- Gestão de portfólios de investimentos
- Predição de preço de ativos financeiros
- Tomada de decisão para investimentos públicos e privados
- Predição de escassez de insumos
- Predição de liquidez de moeda na economia
- Predição de demanda de ativos financeiros e produtos na economia

Estes pontos focam em tornar nossa economia e portanto sociedade mais saudável, impedindo que ocorram problemas para a saúde financeira de empresas, instituições e governos, obtendo assim maior desempenho e autonomia.

REFERÊNCIAS

ACADEMY, D. S. **Deep Learning Book**. 2019. Disponível em: <http://deeplearningbook.com.br∼/introducao-as-redes-neurais-convolucionais/>. Acesso em: 19 de outubro de 2020.

____. **POR QUE A LINGUAGEM PYTHON É TÃO POPULAR EM MACHINE LEARNING E INTELIGÊNCIA ARTIFICIAL?** 2020. Disponível em: <http://datascienceacademy.com.br/∼2blog/por-que-a-linguagem-python-e-tao-popular-em-machine-learning-e-inteligencia-artificial/>. Acesso em: 3 de novembro de 2020.

AIKMAN, D.; HALDANE, A. G.; NELSON, B. D. Curbing the credit cycle. **The Economic Journal**, Oxford University Press Oxford, UK, v. 125, n. 585, p. 1072–1109, 2015.

BAILY, M. N. *et al.* The origins of the financial crisis. Initiative on Business and Public Policy at Brookings, 2008.

BLUWSTEIN, K. *et al.* Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach. Bank of England Working Paper, 2020.

BORIO, C. E.; LOWE, P. W. Asset prices, financial and monetary stability: exploring the nexus. BIS working paper, 2002.

BREIMAN, L. Random forests. **UC Berkeley TR567**, 1999.

BUSSIERE, M.; FRATZSCHER, M. Towards a new early warning system of financial crises. **journal of International Money and Finance**, Elsevier, v. 25, n. 6, p. 953–973, 2006.

CARLANTONIO, L. M. di. **Novas metodologias para clusterização de dados**. 157 p. Dissertação (Dissertação de Mestrado em Ciência da Computação) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001.

CHATZIS, S. P. *et al.* Forecasting stock market crisis events using deep and statistical machine learning techniques. **Expert systems with applications**, Elsevier, v. 112, p. 353–371, 2018.

CHOI, H.; SON, H.; KIM, C. Predicting financial distress of contractors in the construction industry using ensemble learning. **Expert Systems with Applications**, Elsevier, v. 110, p. 1–10, 2018.

COMINCIOLI, B. *et al.* The stock market as a leading economic indicator: An application of granger causality. **The Ames Library**, 1995.

COPPIN, B. **Inteligência Artificial**. LTC, 2010. ISBN 9788521617297. Disponível em: <https://books.google.com.br/books?id=Z3UURQAACAAJ>.

COSTA, C. D. **Top Programming Languages for Data Science in 2020**. 2020. Disponível em: <https://towardsdatascience.com/∼/top-programming-languages-for-data-science-in-2020-3425d756e2a7>. Acesso em 25 de abril de 2021.

DEVELOPERS, E. de suporte G. **Overview of hyperparameter tuning**. 2021. Disponível em: <https://cloud.google.com∼/ai-platform/training/docs/hyperparameter-tuning-overview?hl=pt-br>. Acesso em 30 de abril de 2021.

DREHMANN, M.; BORIO, C. E.; TSATSARONIS, K. Anchoring countercyclical capital buffers: the role of credit aggregates. BIS Working paper, 2011.

FACURE, M. **Semi-Supervised Learning for Fraud Detection**. 2017. Disponível em: <https://lamfo-unb.github.io/∼2017/05/09/Semi-Supervised-learning-for-fraud-detection-Part-1/>. Acesso em: 7 de outubro de 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996.

FISCHER, T.; KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. **European Journal of Operational Research**, Elsevier, v. 270, n. 2, p. 654–669, 2018.

FLAIR, D. **Python SciPy Tutorial – What is SciPy How to Install SciPy**. 2020. Disponível em: <https://data-flair.training/∼ /blogs/scipy-tutorial/>. Acesso em: 11 de novembro de 2020.

FRIEDMAN, J. *et al.* **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. v. 1.

GHOSH, I.; JANA, R. K.; SANYAL, M. K. Analysis of temporal pattern, causal interaction and predictive modeling of financial markets using nonlinear dynamics, econometric models and machine learning algorithms. **Applied Soft Computing**, Elsevier, v. 82, p. 105553, 2019.

GOGAS, P.; PAPADIMITRIOU, T.; AGRAPETIDOU, A. Forecasting bank failures and stress testing: A machine learning approach. **International Journal of Forecasting**, Elsevier, v. 34, n. 3, p. 440–455, 2018.

GONZALEZ, L. d. A. Regressão logística e suas aplicações. Universidade Federal do Maranhão, 2018.

HAO, K. **Three charts show how China's AI industry is propped up by three companies**. 2019. Disponível em: <http://https://www.technologyreview.com/∼2019/01/22/137760/the-future-of-chinas-ai-industry-is-in-the-hands-of-just-three-companies>. Acesso em: 5 de outubro de 2020.

HAYEK, F. A. **Desemprego e Política Monetŕaria**. 2. ed.. ed. São Paulo: Mises Brasil, 2011.

HUANG, Y.-P.; YEN, M.-F. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. **Applied Soft Computing**, Elsevier, v. 83, p. 105663, 2019.

JORDÀ, Ò.; SCHULARICK, M.; TAYLOR, A. M. Macrofinancial history and the new business cycle facts. **NBER macroeconomics annual**, University of Chicago Press Chicago, IL, v. 31, n. 1, p. 213–263, 2017.

JOSHI, A. **We need more Interactive Data Visualization tools (for the Web) in Python**. 2018. Disponível em: <https://medium.com/∼ /@alark/we-need-more-interactive-data-visualization-tools-for-the-web-in-python-ad80ec3f440e>. Acesso em: 11 de novembro de 2020.

KEBANDE, V. R. *et al.* Quantifying the need for supervised machine learning in conducting live forensic analysis of emergent configurations (eco) in iot environments. **Forensic Science International: Reports**, Elsevier, v. 2, p. 100122, 2020.

KRAUSS, C.; DO, X. A.; HUCK, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. **European Journal of Operational Research**, Elsevier, v. 259, n. 2, p. 689–702, 2017.

LEE, K.-F. **Inteligência Artificial: como robôs estão mudando o mundo, a forma como amamos, nos comunicamos e vivemos**. 1. ed.. ed. Rio de Janeiro: Globo Livros, 2019.

LELIS, L. H. S. de. **Aprendizagem Semi-Supervisionada aplicada à Engenharia Financeira**. 128 p. Dissertação (Dissertação de Mestrado em Engenharia Elétrica) — Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

LIEBER, D. *et al.* Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. **Procedia Cirp**, Elsevier, v. 7, p. 193–198, 2013.

LUGER, G. **Inteligência Artificial**. PEARSON BRASIL, 2013. ISBN 9788581435503. Disponível em: <https://books.google.com.br/books?id=UNEKvQEACAAJ>.

MASMOUDI, K.; ABID, L.; MASMOUDI, A. Credit risk modeling using bayesian network with a latent variable. **Expert Systems with Applications**, Elsevier, v. 127, p. 157–166, 2019.

MATOS, D. **Por que Cientistas de Dados escolhem Python?** 2020. Disponível em: <http://www.cienciaedados.com/∼2por-que-cientistas-de-dados-escolhem-python/>. Acesso em: 3 de novembro de 2020.

MEDICAL, S. **Subtle Medical Receives FDA 510(k) Clearance and CE Mark Approval for SubtlePET**. 2018. Disponível em: <http://subtlemedical.com∼/subtle-medical-receives-fda-510k-clearance-and-ce-mark-approval-for-subtlepet/>. Acesso em: 19 de outubro de 2020.

MISES, L. von. **Sobre dinheiro e inflação**. 1. ed.. ed. São Paulo: Vide Editorial, 2017.

MITCHELL, T. **Machine Learning**. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: <https://books.google.com.br/books?id=EoYBngEACAAJ>.

OLIVEIRA, A. F. de. **Favorecendo o Desempenho do k-Means via Métodos de Inicialização de Centroides**. 113 p. Dissertação (Dissertação de Mestrado em Ciência da Computação) — Centro Universitário Campo Limpo Paulista, São Paulo, 2018.

OZBAYOGLU, A. M.; GUDELEK, M. U.; SEZER, O. B. Deep learning for financial applications: A survey. **Applied Soft Computing**, Elsevier, p. 106384, 2020.

PAL, S.; GHOSH, S.; NAG, A. Sentiment analysis in the light of lstm recurrent neural networks. **International Journal of Synthetic Emotions (IJSE)**, IGI Global, v. 9, n. 1, p. 33–39, 2018.

PETROPOULOS, A. *et al.* Predicting bank insolvencies using machine learning techniques. **International Journal of Forecasting**, Elsevier, 2020.

PHI, M. **Illustrated Guide to LSTM's and GRU's: A step by step explanation**. 2018. Disponível em: <https://towardsdatascience.com∼/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>. Acesso em: 19 de outubro de 2020.

PYTHON. **Logo Linguagem Python**. 2020. Disponível em: <https://www.python.org/>. Acesso em: 3 de novembro de 2020.

QU, Y. *et al.* Review of bankruptcy prediction using machine learning and deep learning techniques. **Procedia Computer Science**, Elsevier, v. 162, p. 895–899, 2019.

RODRIGUES, V. **Entenda o que é AUC e ROC nos modelos de Machine Learning**. 2018. Disponível em: <https://medium.com/∼ bio-data-blog/entenda-o-que-%C3%A9-auc-e-roc-nos-modelos-de-machine-learning-8191fb4df772>. Acesso em: 11 de novembro de 2020.

ROTHBARD, M. N. **A grande depressão americana**. [S.l.]: LVM Editora, 2012.

RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 2002.

SAMPAIO, R. F.; MANCINI, M. C. Systematic review studies: a guide for careful synthesis of the scientific evidence. **Brazilian Journal of Physical Therapy**, SciELO Brasil, v. 11, n. 1, p. 83–89, 2007.

SCHULARICK, M.; TAYLOR, A. M. Credit booms gone bust: Monetary policy, leverage cycles, and financial crises, 1870-2008. **American Economic Review**, v. 102, n. 2, p. 1029–61, 2012.

SHUKLA, A. *et al.* Flower classification using supervised learning. **International Journal of Engineering Research Technology**, IJERT, v. 9, 2020.

SOLEYMANI, F.; PAQUET, E. Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder-deepbreath. **Expert Systems with Applications**, Elsevier, p. 113456, 2020.

SOUZA, E. G. de. **Entendendo o que é Matriz de Confusão com Python**. 2019. Disponível em: <https://medium.com/∼ /data-hackers/entendendo-o-que-%C3%A9-matriz-de-confus%C3%A3o-com-python-114e683ec509>. Acesso em: 11 de novembro de 2020.

TAULLI, T. **Introdução à Inteligência Artificial: Uma abordagem não técnica**. 1. ed.. ed. São Paulo: Novatec Editora Ltda, 2020.

TECH, D. **O que é e como funciona o algoritmo RandomForest**. 2020. Disponível em: <https://didatica.tech//∼/o-que-e-e-como-funciona-o-algoritmo-randomforest/>. Acesso em: 13 de outubro de 2020.

WANG, G. *et al.* Financial distress prediction: Regularized sparse-based random subspace with er aggregation rule incorporating textual disclosures. **Applied Soft Computing**, Elsevier, v. 90, p. 106152, 2020.

WANG, P.; ZONG, L.; MA, Y. An integrated early warning system for stock market turbulence. **Expert Systems with Applications**, Elsevier, p. 113463, 2020.