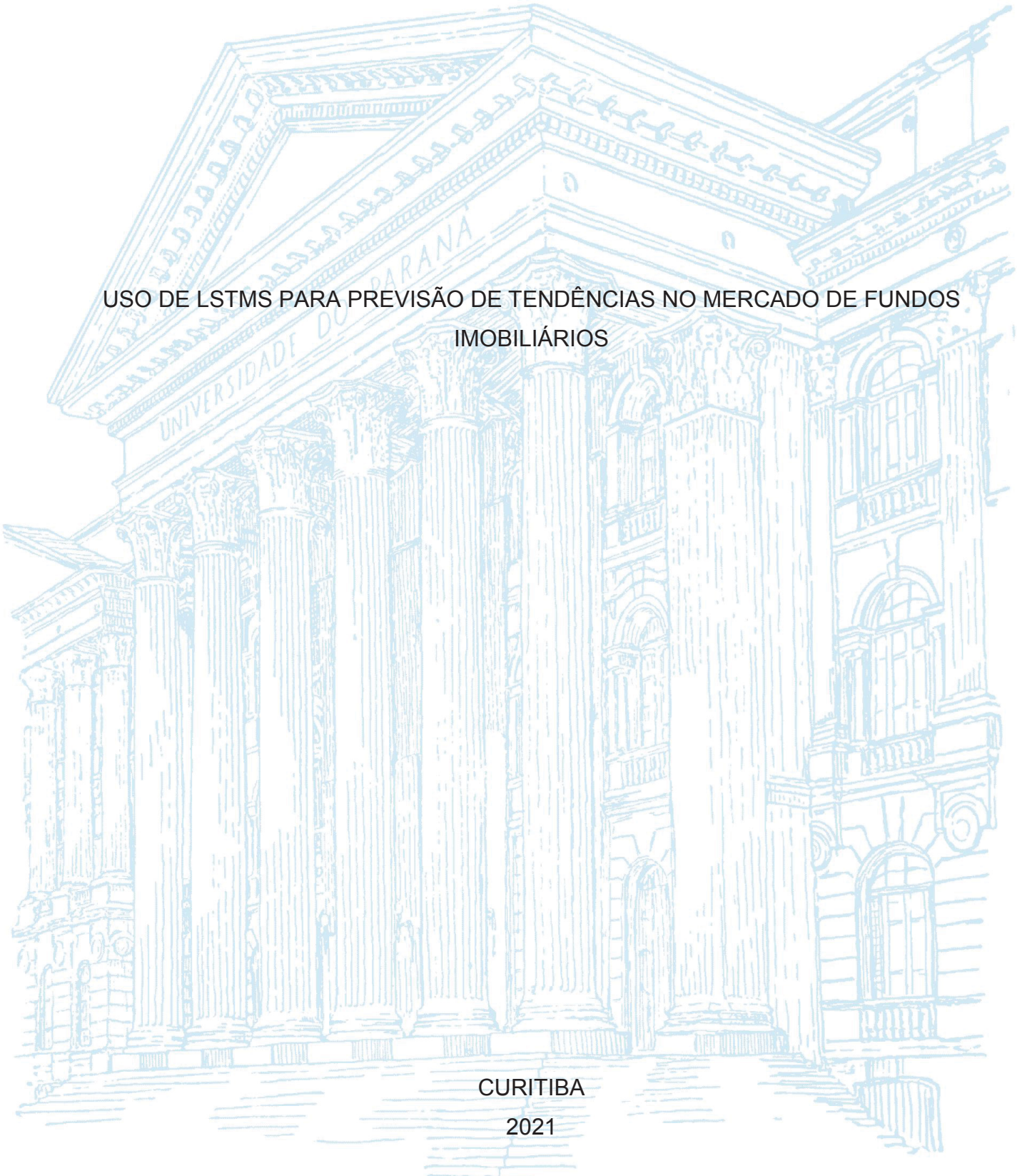
UNIVERSIDADE FEDERAL DO PARANÁ

EVANDRO KONDRAT



USO DE LSTMS PARA PREVISÃO DE TENDÊNCIAS NO MERCADO DE FUNDOS IMOBILIÁRIOS

CURITIBA

2021

EVANDRO KONDRAT

USO DE LSTMS PARA PREVISÃO DE TENDÊNCIAS NO MERCADO DE FUNDOS IMOBILIÁRIOS

Monografia apresentada ao curso de Pós-Graduação em Inteligência Artificial Aplicada, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Especialista em Inteligência Artificial Aplicada.

Orientador: Prof. Dr. Razer Anthom Nizer Rojas Montaño

CURITIBA

2021

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INTELIGÊNCIA ARTIFICIAL APLICADA da Universidade Federal do Paraná foram convocados para realizar a arguição da Monografia de Especialização de **EVANDRO KONDRAT** intitulada: **Uso de LSTMs para previsão de tendências no mercado de fundos imobiliários**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua ____APROVAÇÃO_____ no rito de defesa.
A outorga do título de especialista está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 17 de Junho de 2021.

RAZER ANTHOM NIZER ROJAS MONTAÑO
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

JAIME WOJCIECHOWSKI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Uso de LSTMs para previsão de tendências no mercado de fundos imobiliários

Evandro Kondrat
Especialização em Inteligência Artificial Aplicada
Universidade Federal do Paraná (UFPR)
Curitiba, Brasil
evandrok@gmail.com

Razer Anthom Nizer Rojas Montaño
Especialização em Inteligência Artificial Aplicada
Universidade Federal do Paraná (UFPR)
Curitiba, Brasil
razer@ufpr.br

***Resumo*—O mercado de renda variável está em forte expansão no Brasil. Neste cenário o estudo de séries temporais financeiras é importante para compreensão, previsão de valores futuros e tomada de decisões eficientes nesse mercado. Este trabalho tem por objetivo identificar o potencial de se utilizar redes neurais recorrentes do tipo *Long Short-Term Memory (LSTM)*, para tomada de decisão diária e de forma a obter ganhos consistentes no mercado de fundos imobiliários brasileiro. A partir do uso de séries temporais da variação dos preços desses fundos, índices gerais de mercado, câmbio entre outros.**

***Abstract*—The variable income market is booming in Brazil. In this scenario, the study of financial time series is important for understanding, forecasting future values and making efficient decision in this market. This work aims to identify the potential of using some type of recurrent neural networks called Long Short-Term Memory (LSTM), for daily decision making and in order to obtain consistent gains in the Brazilian real estate market. Using time series of price variation for these funds, general market indices, exchange rates, among others.**

***Index Terms*—ifix, lstm, time series, forecast**

I. Introdução

Dentre as diversas alternativas de investimento em renda variável no Brasil, uma opção que ganhou bastante destaque nos últimos anos é o investimento em Fundos Imobiliários.

Para a escolha sobre seus investimentos, o investidor depende de conhecimento e do acesso à diversas informações sobre as opções existentes. É neste sentido que modelos inteligentes podem atuar para auxiliar o investidor. Essa ideia não é nova, modelos preditivos para o mercado financeiro são tão antigos quanto a própria bolsa de valores [5]. Investidores sempre estão em busca de informações que possam influenciar positivamente seus ganhos. Nesse sentido, diversos estudos foram feitos e podemos separá-los em dois grupos, apresentados a seguir:

O primeiro grupo defende que não é possível aferir ganhos no mercado com a predição de valores futuros, pois, o preço atual dos ativos é o melhor prognóstico sobre o preço futuro. Nessa vertente, temos dois importantes trabalhos: Fama [14] com a hipótese do Mercado Eficiente, onde diz que toda a informação disponível já está refletida nos preços de forma praticamente imediata para mercados suficientemente líquidos; A hipótese do Passeio Aleatório (*random walk*) [24], onde os preços caminhariam de forma aleatória ao longo do tempo, não sendo possível uma predição de comportamentos futuros.

O segundo grupo, defende que os mercados não são realmente eficientes e portanto, seria possível obter retornos acima da média de forma constante explorando as ineficiências do mercado. Este trabalho parte desse pressuposto e vai explorar o mercado de fundos imobiliários brasileiro para tentar encontrar estas ineficiências e fazer previsões acertadas sobre os movimentos futuros de determinados ativos.

A. Motivações

Historicamente o Brasil sempre favoreceu o investimento em renda fixa. Teve uma das maiores taxas de juros básicas do mundo, chegando a 45%a.a. [4]. Assim o investidor conseguia obter altos ganhos com riscos muito baixos.

A queda sistemática e gradual da taxa SELIC, chegando hoje nas mínimas históricas, fez com que até mesmo o mais tradicional investimento brasileiro, a poupança, ter rendimentos próximos a zero [16]. Nesse cenário, mudanças na forma de investir são necessárias, o investidor que quer ter ganhos acima da inflação precisa migrar para investimentos mais arriscados.

Fato corroborado pelos dados do ano de 2020 da B3[1], que mostram que o número de pessoas físicas que investem na bolsa de valores brasileira, cresceu 168% nos últimos dois anos [2].

Diante dessa circunstância, novas ferramentas se fazem necessárias para auxiliar esse alto número de novos investidores com pouca experiência em investimentos de renda variável. É nesse sentido, que o uso de sistemas inteligentes podem tornar-se um grande aliado do investidor.

B. Objetivos

O objetivo desse artigo é analisar se o uso de modelos de inteligência artificial (IA) são capazes de prever a variação dos preços das cotas de fundos imobiliários. Estes preços estão sujeitos à volatilidade do mercado de capitais e da economia, além dos riscos relacionados aos imóveis integrantes dos fundos, como taxa de ocupação e desvalorização dos imóveis.

As análises serão feitas sobre o conjunto dos principais FIIs negociados na bolsa de valores brasileira, utilizando como entrada suas cotações históricas e variáveis do cenário macroeconômico.

[1]B3 é sigla para Brasil, Bolsa, Balcão. A bolsa de valores oficial do Brasil.

Neste contexto, são propostos dois objetivos específicos para este trabalho, a saber:

- Desenvolver modelos de aprendizado de máquina para prever o retorno seguinte (t+1) do preço de cotas de fundos imobiliários;
- Verificar se o uso de indicadores macroeconômicos exógenos ajudam na capacidade preditiva dos modelos.

Para responder os itens citados, serão utilizados modelos redes LSTMs com diversas configurações. Os resultados dos modelos serão comparados com o desempenho das cotas no mesmo período e estratégias de negociação simplificadas.

C. Justificativa

A escolha pelos fundos de Investimento Imobiliário (FII) não é por acaso. Os fatores que motivaram a escolha estão listados a seguir:

- Apesar da lei que regula o funcionamento dos FIIs ter sido promulgada em 1993, a negociação desses fundos só começou a ganhar relevância a partir de 2011;
- Foram encontrados poucos estudos específicos com o foco na previsão de retornos desse tipo de investimento. Entre eles, Carvalho (2019) onde foi usado uma regressão por mínimos quadrados parciais para prever valores futuros do índice IFIX. E dois trabalhos que faziam uso de *wavelets* e redes neurais artificiais para modelos de previsão dos FIIs [32] [13];
- Embora o crescimento do número de pessoas físicas que investem na bolsa de valores brasileira nos últimos dois anos já foi expressivo, o crescimento dos que investem nestes fundos é ainda maior. Mais de 600% nos últimos dois anos [3];
- Apesar de todo esse crescimento, o volume financeiro negociado dos FIIs é cerca de 100 vezes menor que o do mercado acionário. São as pessoas físicas que dominam esse mercado, com quase 70% do volume negociado, ante participação de aproximadamente 25% no mercado de ações, que é dominado por investidores estrangeiros e institucionais.

Conforme o cenário descrito anteriormente, em que o mercado analisado é dominado por pessoas físicas com pouca experiência em renda variável, com relativa baixa liquidez e a falta de estudos específicos, pode ter sua eficiência comprometida, tornando-o assim mais fácil encontrarmos oportunidades de ganhos consistentes.

D. Organização do artigo

Esse artigo é composto por 5 capítulos. No capítulo II de Fundamentação Teórica serão apresentados dois assuntos principais. Primeiro, será feita uma breve introdução sobre mercado financeiro e de fundos imobiliários necessários para a compreensão do trabalho. Em seguida serão apresentados os principais conceitos das técnicas usadas neste trabalho.

O capítulo 3 inicia com uma discussão acerca dos dados a serem utilizados nos experimentos. Após serão detalhados os modelos e arquitetura das Redes Neurais utilizadas. Em seguida, as métricas de como os resultados serão avaliados.

Depois serão demonstrados, comparados e analisados os resultados obtidos pelas diferentes configurações de arquiteturas utilizadas. Por fim, nas considerações finais são discutidos os resultados, limitações e as perspectivas de aplicação da metodologia.

II. Fundamentação Teórica

Neste capítulo são apresentados os conceitos utilizados ao longo desse trabalho.

A. Mercado Financeiro

O objetivo desta seção é esclarecer os termos financeiros utilizados, uma vez que o seu entendimento é essencial para a melhor compreensão do que será apresentado a seguir e nem sempre são conhecidos pelo público em geral [27]. Como os conceitos aqui apresentados já estão bem definidos na literatura, será utilizado como base o que é apresentado em referências consolidadas.

1) Fundos de Investimentos Imobiliários:

Os Fundos de Investimentos Imobiliários (FIIs) são fundos de investimentos destinados à aplicação em empreendimentos imobiliários que incluem, construção de imóveis, aquisição de imóveis prontos, assim como o investimento em títulos relacionados ao mercado imobiliário.

Investimentos em FIIs tem o objetivo de auferir ganhos através de rendimentos de locação ou venda de unidades do empreendimento que lastreia o fundo [9]. Rendimentos que são distribuídos aos cotistas de acordo com o número de cotas que possuir. Além desses, há ganhos com valorização das cotas negociadas em bolsa.

Apesar desses ativos estarem limitados ao mercado imobiliário, podem ter diversas configurações, estratégias e estruturas. Para permitir uma distinção adequada dos fundos, a ANBIMA[2] propõe a seguinte classificação:

- **Desenvolvimento:** fundos que investem majoritariamente, direta ou indiretamente, em desenvolvimento/incorporação de empreendimentos imobiliários em fase de projeto ou construção, para fins de geração de renda com locação ou alienação futura;
- **Renda:** fundos que investem majoritariamente, direta ou indiretamente, em empreendimentos imobiliários construídos, para fins de geração de renda com locação;
- **Títulos e Valores Mobiliários:** fundos que investem majoritariamente, direta ou indiretamente, em títulos de base imobiliária (CRI[3], LH[4] e LCI[5]) e fundos de fundos (cotas de outros fundos imobiliários);
- **Híbrido:** fundos cuja estratégia de investimento não se encaixa nas classificações mencionadas acima.

[2]Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais

[3]Certificado de Recebíveis Imobiliários

[4]Letras Hipotecárias

[5]Letra de Crédito Imobiliário

O patrimônio dos FIIs é dividido em cotas, que representam frações de seu patrimônio. O investidor que deseja investir nesse tipo de aplicação deve, portanto, adquirir estas cotas. É possível adquirir as cotas durante as chamadas ofertas públicas ou negociando com outros investidores no chamado mercado secundário.

As principais vantagens deste tipo de investimento, são as seguintes [9]:

- **Investimento inicial baixo:** A formatação desse investimento em cotas permite ao investidor comprar apenas uma fração de um ou mais imóveis com um valor financeiro inicial mais acessível;
- **Liquidez:** As cotas de FIIs são negociadas em bolsa, como se fossem ações. O que proporciona maior agilidade nas transações em comparação com transações imobiliárias tradicionais;
- **Diversificação:** Um único fundo pode ter uma grande diversidade de imóveis na carteira. A diversificação contribui para diminuir o risco do investidor, pois, torna possível ter ganhos em diferentes cenários e mercados;
- **Gestão profissional:** Os fundos são administrados por gestores profissionais terceirizados, para que os cotistas não precisem se preocupar com procedimentos referentes a certidões, escrituras, recolhimento do ITBI[6], entre outros;
- **Benefício fiscal:** Os rendimentos distribuídos pelos FIIs a seus cotistas são isentos de imposto de renda. Essa é uma das principais vantagens dos fundos imobiliários para a pessoa física. Para haver isenção, o FII deve ter, no mínimo, 50 cotistas. Além disso, o cotista beneficiado pela isenção fiscal não pode deter mais de 10% das cotas do fundo.

As desvantagens desse tipo de investimento estão elencadas pelos próprios regulamentos dos fundos, que citam, em geral: o risco de mercado (e.g. variações do preço do ativo em si), mercado imobiliário (e.g. vacância, desvalorização dos imóveis), da economia (e.g. inflação, taxa de juros) e de gestão (e.g. política de investimento).

O principal indicador do desempenho médio das cotações dos fundos imobiliários negociados nos mercados de bolsa é chamado de Índice de Fundos de Investimentos Imobiliários (IFIX). Ele é o resultado de uma carteira teórica de ativos, elaborada de acordo com certos critérios definidos pela B3. Foi criado no final de 2010 e atualmente sua carteira teórica possui 119 fundos.

Tabela I
OS 5 ATIVOS COM O MAIOR PARTICIPAÇÃO NO IFIX

Fundo	Qtde. Teórica	Participação
KNIP11	47.134.764	5,16%
KNCR11	39.097.087	3,66%
KNRI11	24.157.121	3,48%
HGLG11	18.436.967	3,19%
XPLG11	27.101.061	3,15%

[6]Imposto sobre a Transmissão de Bens Imóveis

2) Compra e venda de ativos:

Existem várias maneiras de operar no mercado financeiro de acordo com o critério e perfil de cada investidor. Essas operações podem ser realizadas com uma alta frequência, baixa frequência, ou apenas uma vez. Cada frequência recebe um nome específico, a saber:

- ***Scalping*** mantêm posições abertas apenas por segundos ou minutos, são operações de alta frequência realizadas sempre através de robôs;
- ***Day trade*** as posições têm que ser abertas e fechadas no mesmo dia;
- ***Swing trading*** mantêm posições por vários dias, semanas ou até alguns meses;
- ***Buy & hold*** são operações feitas por investidores que visam o longo prazo, mantendo a posição aberta por meses ou anos.

Em relação ao tipo de operação em bolsa de valores, existem basicamente duas formas, apresentadas a seguir:

- **Normal:** esta é a forma mais conhecida de negociação. Determinado ativo é comprado por um preço e, após certo período, vendido por outro preço. A diferença entre os preços de compra e venda resultam no lucro ou prejuízo para o investidor. É a operação feita quando se acredita na alta de algum papel;
- **Descoberto:** Neste método, um investidor começa o processo vendendo um ativo, sem efetivamente possuí-lo. Ao fim de determinado período, o investidor compra o ativo anteriormente vendido, fechando a posição. Esta modalidade de negociação permite que investidores lucrem com a queda do preço do ativo. A operação de venda sem possuir os ativos só é permitida em dois cenários: o investidor precisa primeiramente alugar o papel que quer negociar e só então realizar a venda ou no caso da operação de compra ser realizada no mesmo dia (*day trade*), esse aluguel não é necessário.

O empréstimo de FIIs ficou disponível somente a partir de novembro de 2020 [3], antes dessa data não era possível operar a descoberto nesse tipo de ativo.

3) Indicadores macroeconômicos e financeiros:

Os indicadores macroeconômicos são conjuntos de dados e medidas capazes de dar um entendimento do estado da economia em determinado período. Podem ser separados em 4 grandes grupos: indicadores de nível de atividade, inflação, fluxo financeiro e por fim taxas de juros e de câmbio.

Estudos indicam que retornos dos fundos imobiliários podem ser explicados por fatores macroeconômicos (taxa de juros, inflação, economia internacional) [33] principalmente pelo fato dos contratos de locação que lastreiam boa parte dos fundos serem corrigidos por índices de inflação.

Índices de inflação oficiais[7] são mensurados de forma

[7]IPCA (índice nacional de preços ao consumidor amplo) e IGP-M (índice geral de preços - mercado)

mensal, uma alternativa viável que reflete as expectativas da inflação de forma diária podem ser encontradas nos Índices de Mercado ANBIMA, conhecido como IMA, que é referência para os investimentos em renda fixa.

Assim segue uma breve descrição de alguns indicadores:

- **Câmbio** Taxa de conversão de Real para Dólar;
- **EMBI**[8] Índice de Títulos da Dívida de Mercados Emergentes, a principal ferramenta usada pelo mercado financeiro para se chegar ao risco-país;
- **IFIX** principal indicador de desempenho dos fundos imobiliários negociados na B3;
- **IBOV** principal indicador de desempenho das ações negociadas na B3;
- **S&P 500**[9] é um índice americano composto pelas 500 maiores empresas dos EUA. Considerado o índice de maior representatividade no mercado financeiro mundial;
- **VIX**[10] é o indicador que mensura a volatilidade das opções sobre ações do S&P 500. Usado para mediar a incerteza do mercado;
- **IMA-Geral** formado por todos os títulos que compõem a dívida pública;
- **IMA-B** formado por títulos públicos indexados à inflação medida pelo IPCA;
- **IMA-C** formado por títulos públicos indexados à inflação medida pelo IGP-M;
- **IMA-S** formado por títulos pós-fixados atrelados à taxa básica de juros (Selic);
- **IRF-M** formado por títulos públicos prefixados.

4) Análise e Eficiência do Mercado:

A formação de preços de um ativo envolve uma série de fatores econômicos, políticos e até mesmo psicológicos, que podem se relacionam entre si. Entre eles podem-se destacar: a lei da oferta e demanda, perspectivas de crescimento e lucro futuro e especulação financeira.

A análise do mercado e preços segue duas abordagens distintas, que são denominadas análise técnica e análise fundamentalista.

A **análise técnica** utiliza a ideia de que os movimentos de alta ou baixa são determinados por padrões recorrentes, sendo possível prever estes movimentos. Esta análise utiliza-se de gráficos e históricos de preço e volume para identificar padrões que auxiliem na decisão de compra ou venda de um determinado ativo.

A **análise fundamentalista** procura precificar um ativo a partir de seus resultados financeiros e operacionais, como rentabilidade, credibilidade, expectativa de crescimento, etc.

A validade dessas abordagens de análise são criticadas com base na Hipótese do Mercado Eficiente (HME) [14], que preenchidas algumas condições, atesta que seria impossível desenvolver modelos para obter ganhos no mercado e que qualquer modelo desenvolvido possui o mesmo desempenho de tentativas aleatórias.

Existem três níveis descritos de eficiência no mercado [14]:

- **Forma fraca de eficiência:** Afirma que o mercado é eficiente em refletir todas as informações públicas disponíveis. Nesse caso, o histórico de preços dos ativos não teria nenhuma informação útil para previsão de movimentos futuros;
- **Forma semiforte de eficiência:** Sustenta que as novas informações são absorvidas pelo mercado rapidamente. Dessa forma, os investidores não conseguem resultados acima do mercado com informações conhecidas;
- **Forma forte de eficiência:** Neste nível, os preços já refletiriam toda a informação disponível, mesmo aquelas que são de conhecimento público.

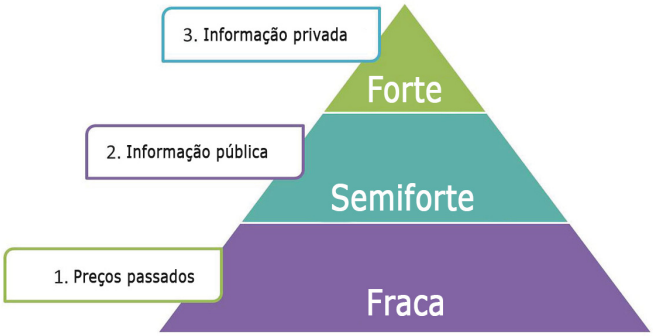


Figura 1. Formas de eficiência, cada nível adiciona um tipo de informação cujo com o qual não seria possível prever um movimento. Fonte: Adaptada de [7].

Apesar de a Hipótese do Mercado Eficiente ser corroborada por vários trabalhos, existem outros que sugerem o contrário, apontando uma série de anomalias e inconsistências não explicadas pela HME [22]. Alguns estudos apontam que os preços não reagem imediatamente à divulgação de novas informações e que levariam um certo tempo para entrarem em equilíbrio. Outra crítica, afirma que a variação dos preços sofrem influência de aspectos psicológicos dos participantes ou que estes podem ser manipulados por grandes investidores.

Tais estudos mostram que os retornos das ações podem ser previstos por apresentarem tendências em determinados períodos e que recursos computacionais têm auxiliado a identificar esses padrões e fazer previsões acertadas.

B. Séries Temporais

Uma série temporal é um conjunto de observações ordenadas no tempo que apresentam uma dependência serial entre as mesmas. Uma grande variedade de fenômenos naturais, econômicos, etc. se enquadram nesse conjunto. Séries temporais podem ter uma grande variedade de padrões e uma maneira comum de descrever o comportamento de uma série

[8]Emergin Markets Bond Index
[9]Standard & Poor's 500
[10]Volatility Index

é através da decomposição da série original em 4 partes: tendência, sazonalidade, ciclo e resíduo [20].

A **tendência** de uma série indica o seu comportamento de longo prazo, isto é, se ela cresce, decresce ou permanece estável, e qual a velocidade destas mudanças.

A **sazonalidade** em uma série corresponde às oscilações de subida e de queda que ocorrem em um determinado período do ano, do mês, etc sempre em frequência conhecida.

Os **ciclos** são caracterizados pelas oscilações de subida e de queda nas séries sem frequência definida. Geralmente ocasionadas por fatores econômicos. A diferença para sazonal é que movimentos cíclicos tendem a ser irregulares e de maior magnitude.

O **resíduo** ou ruído é o que resta após retirar da série os três primeiros componentes. Quando o resíduo é um componente aleatório, de média zero e variância constante, ele recebe o nome de ruído branco (*white noise*). O ruído branco consiste em um tipo de processo puramente aleatório, sem relação causal com nenhum outro fenômeno [10].

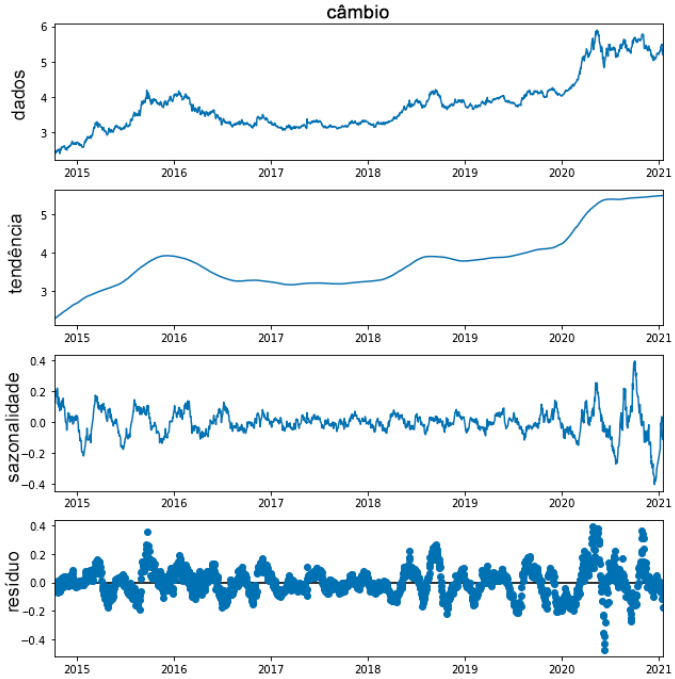


Figura 2. Série temporal dos valores do câmbio, decomposta em tendência, sazonalidade e resíduo. Fonte: Elaborado pelo autor.

Usualmente na decomposição de uma série temporal, os componentes tendência e ciclo são combinados em um único componente, chamado de tendência-ciclo ou apenas tendência [20]. Essa decomposição pode ser escrita como:

$$y_t = T_t \times S_t \times R_t \tag{1}$$

onde y_t é a observação da série no tempo t, T_t é o componente tendência-ciclo, S_t componente da sazonalidade e R_t o resíduo.

Um aspecto fundamental em séries temporais é como uma observação num determinado instante está relacionada às observações passadas, chamado de **autocorrelação**. As observações podem estar autocorrelacionadas em diversas ordens. A autocorrelação de primeira ordem caracteriza séries onde uma observação está correlacionada com a observação imediatamente anterior, segunda ordem quando uma observação está correlacionada com atraso (*lag*) de dois intervalos de tempo e assim por diante.

Outro conceito importante é a **estacionariedade**, que é caracterizada por uma série temporal que se comporta de forma aleatória ao longo do tempo, com média, variância e autocorrelação constantes. Quando uma série possui tendência e/ou sazonalidade ela é considerada como não estacionária e transformações matemáticas podem ser usadas para 'estacionarizar' essa série. A transformação mais comum é fazer diferenças sucessivas da série original, até se obter uma série estacionária. A primeira diferença é definida por [29]:

$$\Delta y_t = y_t - y_{t-1} \tag{2}$$

A figura 3 mostra a diferença de comportamento entre uma série de preços e retornos de um ativo. Uma série de preços costuma ser do tipo não estacionário. Enquanto a série de retornos é representada por um processo estacionário.

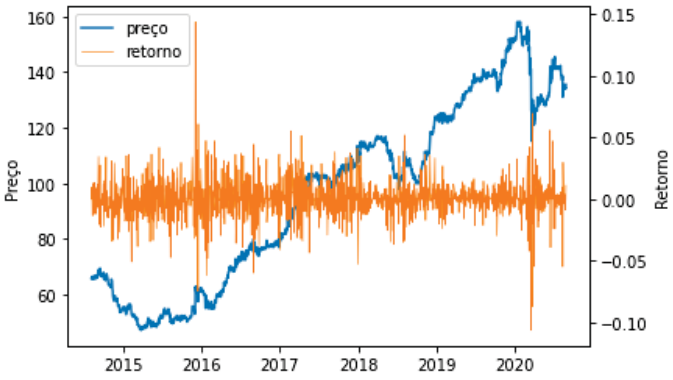


Figura 3. Comparação de séries temporais de preços e retornos de THRA11. Fonte: Elaborado pelo autor.

Em geral os estudos de uma série temporal estão interessados em: Analisar e descrever a série, verificar suas características mais relevantes e suas possíveis relações com outras séries; Previsão (*forecast*), a partir de valores históricos da série, procura-se estimar os valores futuros.

A maioria dos métodos de previsão costuma usar séries temporais estacionárias para fazer previsões por serem séries mais bem comportadas e terem propriedades bem conhecidas. Como, por exemplo, a reversão à média, "um processo estacionário não pode se distanciar demais de sua média, por ter variância finita. A velocidade de reversão à média é determinada pela autocovariância: ela é rápida quando as autocovariâncias são pequenas, e lenta quando são grandes"[28, p. 21].

1) Testes de raiz unitária:

A série de retornos da figura 3 oscila em torno de valores bem definidos e limitados, um sinal claro de estacionariedade. Mas nem toda séria tem um comportamento tão evidente e por isso, é necessário um método objetivo para fazer tal afirmação.

Assim, para determinar estacionariedade em uma série temporal, foram desenvolvidos testes estatísticos chamados de testes de raiz unitária que têm as seguintes hipóteses:

H_0(hipótese nula): tem raiz unitária (não é estacionária)

H_1: não tem raiz unitária (é estacionária)

O primeiro teste foi feito por Dickey e Fuller em 1979 e foi chamado de Dickey-Fuller. Uma de suas premissas é de que os seus resíduos se comportem como ruídos brancos. Contudo, em muitas séries temporais o erro faz parte de um processo estacionário qualquer [11]. Por este motivo, posteriormente foi feito uma melhoria no teste original e esse novo teste recebeu o nome de Dickey-Fuller Aumentado (ADF)[11].

O teste de ADF testa a hipótese nula, na qual, uma raiz unitária está presente em um modelo autorregressivo. Isto é feito decompondo a série em componentes, sendo um deles o coeficiente de presença de raiz unitária. Desse coeficiente é calculado uma estatística, que os valores críticos foram tabelados por Dickey e Fuller através de simulação Monte Carlo [11]. Ao comparar o valor obtido com o tabelado obtém-se a probabilidade de rejeição da hipótese nula.

Existem outros testes de raiz unitária, como o teste de Phillips-Perron e o teste KPSS, mas o ADF que é mais conhecido e presente na literatura [28].

C. Aprendizado de Máquina

O Aprendizado de Máquina (*machine learning*), é uma subárea da Inteligência Artificial, que se utiliza de algoritmos específicos que podem aprender sem serem especificamente programados, através da observação de dados. Apesar de não ser uma área nova, vem ganhando maior relevância nos últimos anos devido a maior disponibilidade de volume de dados e melhoria na capacidade de recursos computacionais [18].

A utilização de aprendizado de máquina destaca-se no processo de obtenção de informação útil a partir de um grande volume de dados onde seria inviável realizar a tarefa manualmente. Essa informação é representada em um tipo de modelo matemático que é usado para fazer inferências, ou predições, a partir de um novo conjunto de dados.

O mecanismo de aprendizagem se divide em duas categorias principais, o aprendizado supervisionado e o aprendizado não supervisionado.

1) Aprendizado Supervisionado:

No aprendizado supervisionado, apresenta-se ao algoritmo exemplos de entradas e saídas desejadas. O objetivo é aprender uma regra geral que mapeia as entradas para as saídas ainda não conhecidas. Os problemas mais comuns desse tipo são os que envolvem classificação ou regressão dos dados.

[11] Augmented Dickey-Fuller

Nos problemas de **classificação**, as entradas são mapeadas de forma a rotular os dados em categorias limitadas e distintas. Um problema frequente, acontece quando o conjunto de dados utilizado para o treinamento do modelo possui muitos exemplos de uma classe e poucos exemplos da outra classe. Esse desbalanceamento de classes pode ser corrigido com técnicas para aumentar ou diminuir as amostras já existentes no conjunto, de forma a se obter um conjunto de dados balanceado.

Nos problemas que envolvem **regressão**, as entradas são mapeadas para uma função continua. Sua principal diferença para a classificação é saída do modelo, que nesse caso é um valor numérico.

Alguns dos algoritmos mais importantes dessa categoria são [18]:

- Regressão Linear;
- Regressão Logística;
- K - *Nearest Neighbors* (KNN);
- Máquinas de vetores de suporte (SVM - *Support Vector Machine*);
- Árvores de decisão;
- Rede neural artificial.

2) Aprendizado Não Supervisionado:

No aprendizado não supervisionado não é apresentado ao algoritmo as saídas desejadas. Nesse caso, o algoritmo sozinho deve encontrar estruturas ou padrões para as entradas fornecidas. Dos resultados é feita uma medição da qualidade da tarefa. Assim, os parâmetros são ajustados de modo que essa medida seja otimizada.

As técnicas de aprendizado de máquina não supervisionadas são aplicadas em problemas de clusterização, associação, detecção de anomalias e redução de dimensionalidade [18].

3) Generalização:

Quando treinamos um modelo de Aprendizado de máquina, a ideia é que o modelo aprenda sobre os dados de entrada e possa realizar previsões, com um erro aceitável, com novos dados, que não estavam presentes nos dados de entrada, ou seja, dados nunca vistos pelo modelo.

Overfitting acontece quando um modelo tem um desempenho excelente com os dados nos quais foi treinado, porém se mostra ineficaz para prever resultados com novos dados. Neste caso, o modelo não tem capacidade de generalização. Aprendeu tão bem as relações existentes no treino, que acabou apenas decorando o que deveria ser feito não sabendo trabalhar com novas amostras.

O *overfitting* acontece quando o modelo é muito complexo em relação a quantidade e ao ruído de dados de treinamento ou devido ao excesso de treinamento. As possíveis solução são [18]:

- Simplificar o modelo escolhendo um algoritmo mais simples, com menos atributos;

- Coletar mais dados para treinamento do modelo;
- Reduzir o ruído nos dados de treinamento, corrigindo ou removendo valores extremos e incorretos;
- Limitar o treinamento de modo a evitar que os dados sejam decorados.

Por outro lado, no ***underfitting*** o modelo não consegue encontrar relações entre as variáveis, neste cenário o desempenho do modelo já é ruim no próprio treinamento.

As causas do *underfitting* podem incluir, o uso de modelos muito simples para os dados, assim ele não é capaz de encontrar relações entre as variáveis. Outra causa comum é a utilização de características não representativas para o modelo, nesse caso o modelo não encontra relações entre as variáveis porque elas simplesmente não existem. As possíveis solução são [18]:

- Selecionar um modelo mais poderoso, com mais parâmetros;
- Escolher melhores características para alimentar o algoritmo de aprendizagem;
- Reduzir as restrições do modelo, tornando-o mais flexível.

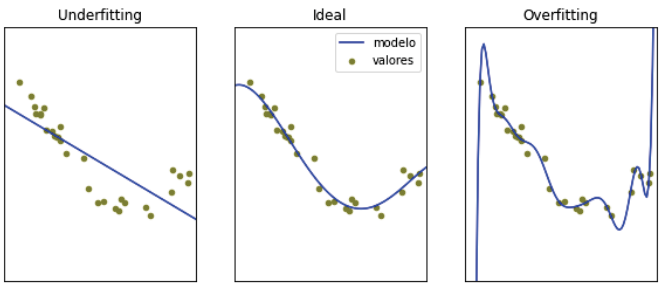


Figura 4. Gráfico demonstrando modelos com diferentes graus de ajuste aos dados de treinamento. Fonte: Elaborado pelo autor.

4) Separação dos dados:

Poderíamos utilizar a totalidade dos dados no processo de treinamento, porém, desta forma não saberíamos o real desempenho deste modelo. Como visto anteriormente, os modelos podem sofrer de *overfitting* que só seriam descobertos na aplicação real do modelo.

Desta forma, para medir o desempenho real do modelo criado, é necessário realizar testes com ele, utilizando dados diferentes dos usados durante o treinamento. Para este fim, a melhor opção é separar os dados em dois conjuntos: conjunto de treinamento e conjunto de testes [18].

Ao realizar o treinamento do modelo usando o conjunto de treinamento e avaliando o mesmo com o conjunto de testes, obtêm-se uma boa estimativa de erro do modelo para um conjunto inédito de dados.

Um bom modelo deve ser capaz de aprender a generalizar, tendo desempenhos semelhantes com os dois conjuntos. Se o erro medido no conjunto de treinamento for muito menor que no de testes significa que o modelo sofreu *overfitting* aos dados.

Existe outro problema relacionado a escolha dos hiperparâmetros. Ao medir o erro por várias vezes no conjunto de testes, existe o risco do modelo e dos hiperparâmetros se adaptarem a este conjunto em específico, e que provavelmente não funcionariam tão bem sob um novo conjunto de dados [18].

Para este problema, uma solução comum é dividir o conjunto de testes em dois, criando um conjunto chamado de validação. Assim o modelo e os hiperparâmetros são selecionados com base no erro do conjunto de validação e testados efetivamente uma única vez no conjunto de testes.

D. Redes Neurais Artificiais

As Redes neurais artificiais (RNA) são um modelo computacional inspirado no funcionamento do cérebro. As primeiras referências desse modelo datam de 1943 com McCulloch e Pitts que modelaram uma rede neural simples usando circuitos elétricos. O processo de desenvolvimento das RNAs é marcado por longos períodos de estagnação e por um recente e crescente interesse científico, devido às inovações tecnológicas e aumento da capacidade computacional. Elas são usadas em modelos de inteligência artificial por possuírem capacidade de aprendizado de máquina e reconhecimento de padrões [15].

Os neurônios artificiais, são chamados de perceptrons que são as estruturas mais básicas da rede. Os perceptrons mais simples possuem três estruturas lógicas: conjunto de entradas cada uma caracterizada por um peso, um somador para somar os sinais de entrada ponderadamente e uma função de ativação que determinará a ativação ou não do perceptron, tipicamente em um intervalo normalizado. O modelo pode incluir também um *bias* de entrada. Ele é incluído no somatório com o objetivo de aumentar o grau de liberdade desta função [1].

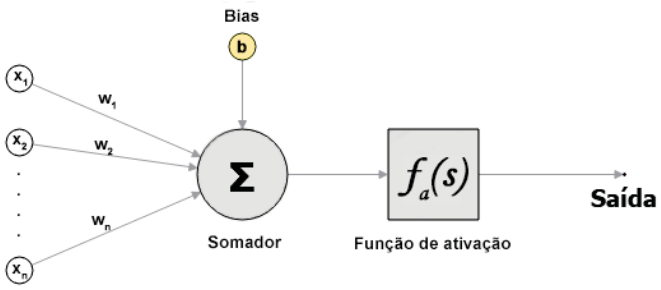


Figura 5. Neurônio artificial recebendo n sinais de entrada mais bias. Fonte: Elaborado pelo autor.

A descrição matemática de um perceptron pode ser vista como:

$$saida = f_a(\sum(entradas \cdot pesos) + bias) \qquad (3)$$

São nestas estruturas que o processo iterativo de ajuste dos pesos acontece, a chamada aprendizagem. A definição de aprendizagem no contexto de redes neurais segundo Mendel e McLaren 1970 é:

"Aprendizado é um processo pelo qual, os parâmetros de uma rede neural são adaptados através de um processo

contínuo de estimulação do ambiente no qual a rede está inserida. O tipo de aprendizado é determinado pela maneira na qual ocorre as mudanças dos parâmetros"

1) Treinamento da rede neural:

Os pesos e bias dos neurônios devem ser estimados da melhor forma possível, de modo a minimizar o erro entre a saída calculada e a saída real. Esse processo de estimação e ajuste é dado o nome de **treinamento da rede neural** [15].

O treinamento começa, na maior parte das vezes, com os pesos e bias inicializados de forma aleatória. Então, segue a seguinte iteração:

- Cálculo da saída com os pesos e bias atuais.
- Cálculo do erro de saída, comparando a saída calculada com a saída real.
- Teste de convergência do erro calculado. É adotado um critério de parada dependendo da análise do problema.
- Aplicação de uma função matemática para atualizar os pesos e bias de forma a minimizar o erro calculado.

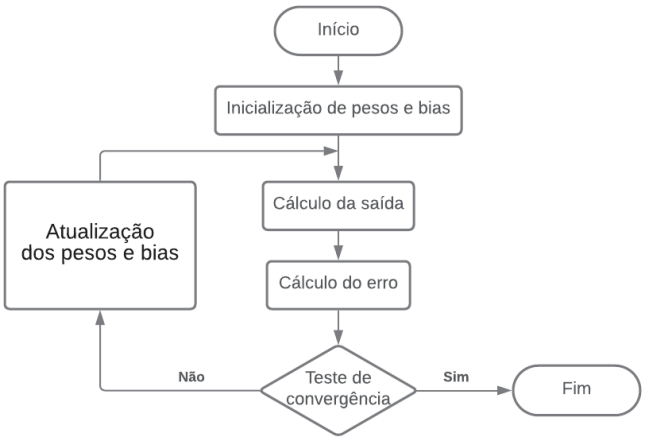


Figura 6. Fluxograma do treinamento de rede neural. Fonte: Elaborado pelo autor.

Cada execução do laço é chamada de **época de treinamento** e pode ser melhor visualizado no fluxograma da figura 6.

2) Camadas:

Tipicamente as redes utilizam conjuntos de neurônios interconectados em várias camadas, chamados de *Multi-Layer Perceptron* (MLP), como vistos na figura 7.

As camadas de um MLP podem ser divididas em três tipos:

- **Entrada** (*input layer*): Cada rede tem uma camada desse tipo. O número de neurônios é igual ao número de atributos dos dados. Em algumas configurações possuem um neurônio adicional para o bias.
- **Escondida** (*hidden layer*): Responsável pelo processamento dos dados. A arquitetura dessa camada pode variar bastante, podendo ir de zero a muitas camadas ocultas. O número de neurônios também varia conforme o tipo e complexidade do problema.
- **Saída** (*output layer*): Igualmente a camada de entrada, cada rede tem apenas uma camada de saída. O número de neurônios é determinado pelo tipo de problema. Em regressões teremos um único nó. Para problemas de classificação, geralmente teremos um nó por classe.

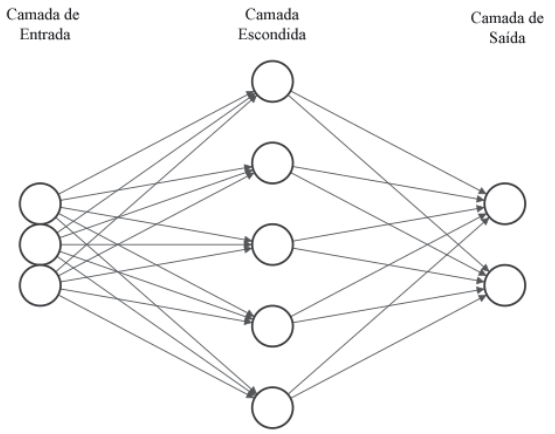


Figura 7. Exemplo de MLP com uma camada escondida de 5 neurônios. Fonte: Elaborado pelo autor.

A topologia mais comum é denominada *feedforward*, onde o fluxo de dados é unidirecional, sem ciclos ou repetições entre seus neurônios [25].

3) Funções de ativação:

O uso de funções de ativação em redes neurais tem algumas finalidades. A principal delas é adicionar um elemento não linear à rede, sem esse elemento, a rede neural, só poderia captar relações lineares entre as variáveis de entrada e de saída [19]. Essa finalidade é mais importante nas camadas escondidas da rede neural.

Nas camadas de saída, as funções de ativação tem a tarefa de transformar e/ou limitar o valor de saída, a depender do tipo de problema que a rede está tentando resolver.

Algumas das funções de ativação mais utilizadas:

- **Função linear:** É a função de ativação mais básica porque não altera a saída de um neurônio.

$$f(s) = s \tag{4}$$

- **Função de Etapa Binária:** Ela pode ser usada quando a saída da rede é um classificador binário.

$$f(s) = \begin{cases} 1, & \text{se } s >= 0 \\ 0, & \text{se } s < 0 \end{cases} \tag{5}$$

- **Função sigmoide:** Nesse tipo de função, a saída do neurônio assumirá valores reais entre 0 e 1.

$$f(s) = \frac{1}{1 + e^{-\beta s}} \tag{6}$$

- **Função tangente hiperbólica:** Nesse tipo de função, a saída do neurônio assumirá valores reais negativos e positivos entre -1 a 1.

$$f(s) = tanh(s) = \frac{1 - e^{-s}}{1 + e^{-s}} \tag{7}$$

- **Função ReLU:** Por ser leve computacionalmente e acelerar a velocidade do treinamento, é hoje uma das funções de ativação mais utilizadas.

$$f(s) = max(0, s) \tag{8}$$

4) Função de Custo:

Para cada iteração de treinamento, é preciso quantificar o erro entre a saída calculada e a saída real. Isso é feito pelas chamadas funções de custo, também chamadas de funções de perda (*loss function*). É uma função de suma importância, sua tarefa é traduzir em um único número, o quão perto do objetivo o modelo está, de modo que uma melhoria desse número indique um modelo melhor [30].

A função mais comum para problemas de regressão é a função de perda de erro quadrático médio (*mean squared erro* - MSE). Dada pela equação 9, ela calcula a média dos erros do modelo ao quadrado, fazendo com que o erro aumente exponencialmente conforme aumenta a distância entre o valor predito e o valor real.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2 \tag{9}$$

Outra função bastante conhecida é chamada de erro absoluto médio (*mean absolute error* – MAE). A diferença para a MSE é que ela calcula a média dos erros de maneira linear, penalizando menos os valores discrepantes (*outliers*).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - X_i| \tag{10}$$

5) Função de Otimização :

O objetivo da função de otimização é, dado uma função arbitrária, encontrar o conjunto de variáveis de entrada que resulte no menor valor da função.

O algoritmo de otimização mais usado para treinamento de redes neurais é conhecido como Descida do Gradiente (*Gradient Descent*), um dos requisitos para seu uso é que a função de custo seja diferenciável. Sua equação é apresentada em 11, onde W são os coeficientes da função, α a taxa de aprendizagem e J a função de perda.

$$W_{j+1} = W_j - \alpha\nabla J(W_j) \tag{11}$$

A derivada do custo (gradiente) é calculada, pois, resulta na inclinação da função em um determinado ponto. Conhecer a inclinação é fundamental para definir a direção (sinal) para atualizar os valores dos coeficientes para obter um custo menor na próxima iteração.

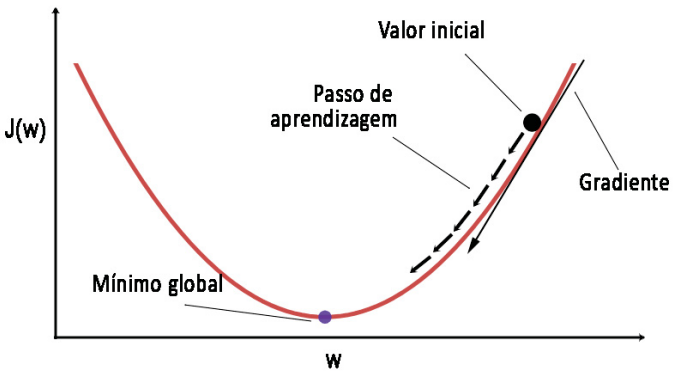


Figura 8. Representação da Descida do Gradiente (com o objetivo de minimizar a função de custo) Fonte: Elaborado pelo autor.

A taxa de aprendizagem controla o quanto os coeficientes alteram em direção oposta ao gradiente em cada atualização. Se essa taxa de aprendizagem for muito pequena, o algoritmo precisará de muitas iterações para convergir, o que levará muito tempo. Por outro lado, se ela for muito grande, o algoritmo pode nunca convergir.

O algoritmo inicializa em um ponto aleatório do espaço de pesos e então, os atualiza na direção oposta ao do gradiente até que algum critério de parada seja atingido.

A cada iteração do algoritmo pode receber um ou mais exemplos antes de fazer uma atualização de peso, essa quantidade é chamada de **tamanho do lote** (*batch size*). Para um lote grande maior será o tempo necessário para calcular o gradiente para cada etapa. Quanto menor o lote, mais irregular será a descida do gradiente. Se um lote contiver todos os exemplos de treinamento então cada iteração é igual a uma época [1].

E. Redes Neurais Recorrentes

Em uma rede neural tradicional, todas as entradas (e saídas) são independentes umas das outras. Mas algumas informações só fazem sentido pelo contexto em que estão inseridas e foi para suprir essa necessidade que as redes neurais recorrentes (RNR) foram criadas. Elas são um tipo de RNA projetada para sequências de dados. Sua diferença, é possuir uma recorrência, onde o resultado do passo anterior (h_{t-1}) é realimentado na rede, afetando assim o passo seguinte (h_t). Assim, as redes recorrentes têm duas fontes de entrada, resultado anterior e o valor de entrada presente (X_t), tal que:

$$h_t = f(h_{t-1}, X_t) \tag{12}$$

onde h_t além da saída é chamado de estado do sistema. A figura 9 representa o conceito recorrente da equação 12, onde o bloco A representa a função $f(.)$.

A função pode ser desdobrada em um número finito de janelas temporais (*time steps*), aplicando a equação 12 por $t - 1$ vezes. Esse desdobramento pode ser visto na figura 10 que ilustra uma rede recorrente com 3 *time steps*.

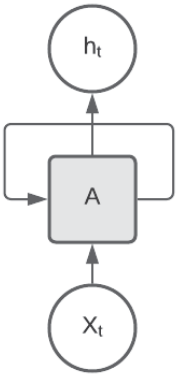


Figura 9. Neurônio recorrente Fonte: Elaborado pelo autor.

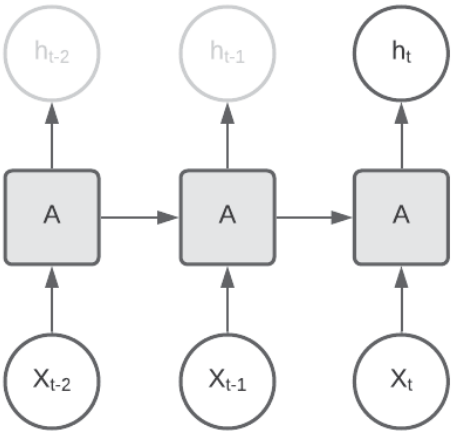


Figura 10. RNR desdobrada, com 3 *time steps*. Fonte: Elaborado pelo autor.

Por essa característica de conseguirem guardar contexto, podem-se dizer que as RNRs têm memória de curto prazo. Naturalmente são aplicadas em problemas onde o contexto é importante, como tradução de textos, caligrafia, reconhecimento de voz, dados de séries numéricas, bolsas de valores, etc.

Uma de suas deficiências é chamada de explosão/dissipação do gradiente. Esse problema aparece em dependências longas, onde as sucessivas operações fazem com que o valor do gradiente tenha a tendência de explodir ou a desaparecer nas camadas anteriores, o que pode gerar problemas na modelagem dos dados.

Possuem diversas variações, as mais conhecidas são:

- Hopfield;
- Gated Recurrent Unit (GRU);
- Long Short Term Memory(LSTM).

F. Redes Long Short Term Memory

As redes *Long Short Term Memory* (LSTM) foram criadas para resolver problemas onde o contexto de longo prazo são importantes. Sua principal diferença para as RNRs é a implementação de um estado da célula que é perpetuado aos passos seguintes. Para adicionar ou remover informações ao estado da célula a LSTM usa estruturas chamadas de portões (*Gates*). Os portões são funções do tipo sigmoide que produzem números entre zero e um, que determina o quanto de cada valor é liberado.

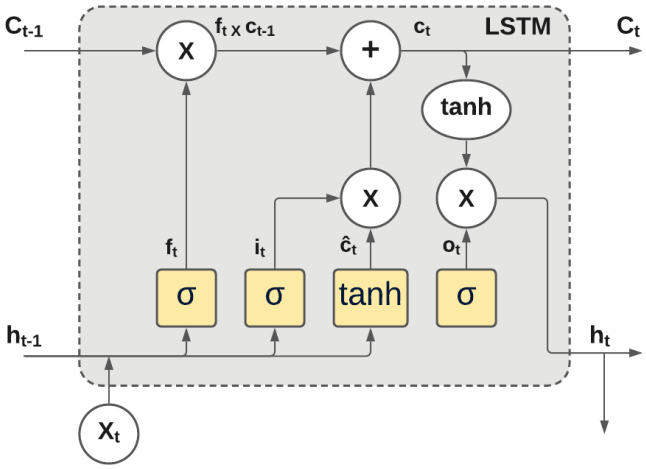


Figura 11. Diagrama de uma célula LSTM com suas estruturas internas. Fonte: Elaborado pelo autor.

O funcionamento de uma LSTM pode ser dividida em quatro estruturas básicas, detalhadas a seguir.

Estado da célula (*Cell State*) é o que da capacidade de a rede de manter informações de longo prazo fora do fluxo convencional. Representada na linha horizontal na parte de cima do diagrama da figura 11.

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{13}$$

Forget Gate o primeiro passo é decidir se o estado anterior da célula será mantido ou esquecido. Essa decisão é feita na *Forget Gate*. Ela tem uma função de ativação sigmoide que recebe os valores h_{t-1} e X_t e retorna um número entre 0 e 1. Um representa manter completamente o estado da célula e zero esquecer completamente.

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_t) \tag{14}$$

Input Gate o passo seguinte é decidir quais informações serão adicionadas no estado da célula. Primeiro, temos outra função sigmoide com as mesmas entradas da *forget Gate* para decidir quais valores atualizar. Esses valores são combinados com um vetor chamado modulação de entrada (16) para atualizar o estado.

Essa modulação de entrada é necessária, uma vez que esses valores são adicionados ao estado da cédula e como a sigmoide retorna apenas números entre 0 e 1, o estado nunca poderia ser zerado/esquecido. Por isso usa-se a função tangente hiperbólica que tem uma saída mais ampla entre -1 e 1.

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \tag{15}$$

$$\tilde{C}_t = tanh(W_c[h_{t-1}, X_t] + b_c) \quad (16)$$

Output Gate e por fim é preciso definir a saída. Os valores de entrada e o estado da célula são combinados após passarem por uma sigmoide e uma função tangente hiperbólica. O resultado é enviado como saída e entrada para próxima célula.

$$o_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \quad (17)$$

$$h_t = o_t \times tanh(C_t) \quad (18)$$

Dessa forma cada célula LSTM recebe três entradas, o último estado da cédula (C_{t-1}), a saída do último estado oculto (h_{t-1}) e a observação no tempo = t (X_t).

III. Materiais e Métodos

Nesse estudo, foi utilizada a série histórica de preços individuais dos fundos que compõem o IFIX. Dos 119 fundos da carteira atual, nem todos possuem um longo histórico, outros tiveram no passado, um volume negociado tão baixo que passavam semanas sem alterações no preço. Por isso a data do estudo foi escolhida de forma que incluísse o maior número possível de fundos, entre agosto de 2014 até abril de 2021.

Nas sessões a seguir serão apresentados com mais detalhes os processos de pré-processamento e separação dos dados. Depois, treinamento e a forma de avaliação dos modelos adotada.

A. Preparação do Conjunto de Dados

Foram coletados a série histórica diária de preços de fechamento dos 119 fundos que compõem o IFIX mais os valores diários dos 11 indicadores vistos na seção II-A3, com 1558 amostras cada um.

Tabela II
Amostra dos dados coletados

Data	...	KNCR11	KNRI11	...	VIX	S&P	CÂMBIO	IFIX	IBOV
11/08/2014	...	103,38	119,50	...	14,23	1.936,92	2,29	1.391,82	56.613,32
12/08/2014	...	103,99	117,90	...	14,13	1.933,75	2,29	1.389,80	56.442,34
13/08/2014	...	104,50	116,12	...	12,90	1.946,72	2,29	1.392,70	55.581,19
14/08/2014	...	103,50	117,00	...	12,42	1.955,18	2,28	1.391,48	55.780,41
15/08/2014	...	106,40	118,44	...	13,15	1.955,06	2,27	1.397,15	56.963,65
...	...	...	...	...	...	...	...	...	...
12/04/2021	...	93,26	151,00	...	16,91	4.127,99	5,74	2.845,01	118.811,74
13/04/2021	...	93,30	150,89	...	16,65	4.141,59	5,72	2.843,37	119.297,13

1) Pré processamento:

Uma inspeção visual constatou que pontualmente existiam para certos dias, valores discrepantes dos restantes, possivelmente uma falha na coleta dos dados. Então a primeira etapa foi a remoção dos valores claramente atípicos dos preços. Para essa etapa, foram removidos todos os valores que tinham uma variação de ± 20% em relação ao dia anterior.

Qualquer série que apresentava mais que 5 valores faltantes consecutivos foram descartadas. Nas demais, foi aplicado uma interpolação linear para preencher esses valores. Também foram descartadas as séries nas quais os valores não apresentaram mudanças por mais de 10 dias consecutivos.

Com estas regras, os problemas apresentados no início desse capítulo foram mitigados, 86 fundos acabaram sendo descartados, restando assim 33 fundos com os 11 indicadores financeiros e macroeconômicos.

2) Retornos e estacionariedade:

Como o objetivo é prever os retornos futuros, foi preciso transformar as séries de preços em séries de retornos diários. Seja P_t o preço de um fundo ou valor de um indicador em determinado dia, seu retorno é definido por [31]:

$$R_t = \frac{P_t}{P_{t-1}} - 1 \quad (19)$$

Tabela III
Resultado do teste de ADF sobre as séries de retornos dos fundos e indicadores

Séries	Estatística ADF[1]	Valor-p[2]
ABCP11	-11,827	0,000
BBFI11B	-21,791	0,000
BBPO11	-13,783	0,000
BBRC11	-47,489	0,000
BPFF11	-17,530	0,000
BRCR11	-17,327	0,000
CBOP11	-18,724	0,000
CEOC11	-12,313	0,000
CTXT11	-7,155	0,000
FAMB11B	-24,805	0,000
FCFL11	-33,857	0,000
FIGS11	-9,112	0,000
FLMA11	-17,664	0,000
FPAB11	-12,360	0,000
HGBS11	-16,941	0,000
HGCR11	-19,711	0,000
HGRE11	-17,004	0,000
JSRE11	-12,744	0,000
KNCR11	-11,683	0,000
KNRI11	-14,071	0,000
MBRF11	-48,164	0,000
MXRF11	-16,816	0,000
PQDP11	-18,256	0,000
RBBV11	-30,541	0,000
RBRD11	-9,672	0,000
RNGO11	-17,137	0,000
SDIL11	-17,789	0,000
SHPH11	-12,001	0,000
SPTW11	-17,592	0,000
VLOL11	-21,293	0,000
VRTA11	-17,205	0,000
XPCM11	-10,043	0,000
XTED11	-14,140	0,000
IMA-B	-41,264	0,000
IMA-C	-40,177	0,000
IMA-S	**-1,039**	**0,739**
IRF-M	-25,355	0,000
IMA-G	-41,024	0,000
EMBI	-17,680	0,000
VIX	-18,378	0,000
S&P	-9,428	0,000
Câmbio	-43,702	0,000
IFIX	-10,220	0,000
IBOV	-15,002	0,000

[1]A estatística ADF, é um número negativo, quanto mais negativo, mais indicativo o teste se torna de rejeitar a hipótese nula de que existe raiz unitária na série. [2]É a probabilidade de rejeição da hipótese nula quando ela é verdadeira.

Sobre as séries de retornos, foi aplicado o teste de ADF para confirmar a estacionariedade em cada uma das séries.

Como já era esperado, quase todas as séries passaram no teste, com uma probabilidade menor que 0,1% de serem não estacionárias, como demostrado na tabela III. Apenas a série de retornos do indicador 'IMA-S' apresentou alta probabilidade de não estacionariedade no teste.

Analisando o indicador, foi visto que como ele se comporta de forma diferente dos demais, apresenta pequenos retornos diários sempre positivos. Por isso, foi optado por descartar esse indicador.

Por fim, a base de dados final ficou com 33 séries de retornos diários dos fundos e 10 séries de indicadores, cada uma com 1558 amostras.

3) Conjunto de treino, validação e teste:

A separação dos conjuntos de treino, validação e teste foram feitas nas proporções aproximadas de: 70%, 15% e 15% respectivamente.

É possível verificar pela figura 12, que o conjunto de dados abrangeu todo o período da pandemia mundial do Coronavírus (COVID-19), fato com consequências inéditas na volatilidade das séries de retornos. Por isso, inicialmente foi excluído um período de 135 dias do conjunto de dados, uma vez que com uma volatilidade tão alta, os dados de um único dia poderiam distorcer a média de todo o período de testes. Ao final dos experimentos iniciais, esse período será incluído para comparação.

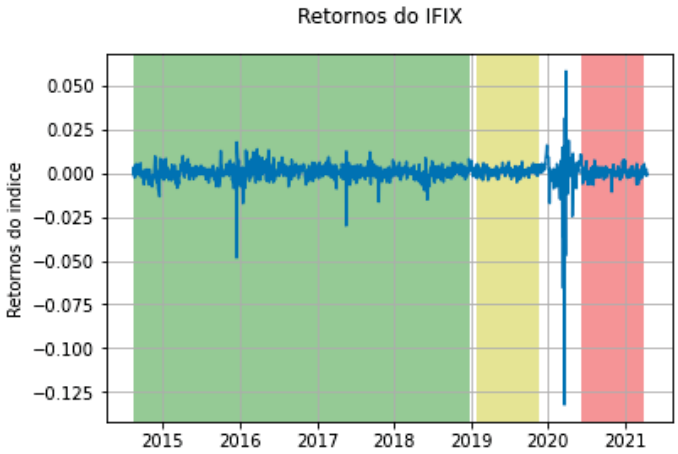


Figura 12. Separação dos dados em treino, validação e teste indicados pelas cores verde, amarela e vermelha respectivamente. Fonte: Elaborado pelo autor.

Assim, o conjunto de treino ficou com 1000 amostras e o conjunto de validação e teste com 200 amostras cada um. Tratando-se séries temporais, os dados não foram embaralhados e foram tomados de forma sequencial, como mostrado no gráfico da figura 12. Os dados não precisaram ser normalizados uma vez que a série de retornos já apresenta a mesma ordem de grandeza para todos os dados.

4) Entrada de dados:

Como um dos objetivos específicos deste artigo é verificar se o uso de indicadores macroeconômicos melhora sua capacidade preditiva das RNAs as entradas foram separadas em 3 conjuntos de dados.

1) Fundo analisado ;
2) Fundo analisado, IMA-G, IFIX, IBOV;
3) Fundo analisado, IMA-G, IFIX, IBOV, Câmbio, EMBI, IMA-B, IMA-C, IRF-M, VIX, S&P.

Como exemplificados na figura 13, o 1º conjunto apenas com o retorno do fundo (laranja). Um segundo que adicionou 3 indicadores (azul) e terceiro conjunto com todos os 10 indicadores (vermelho).

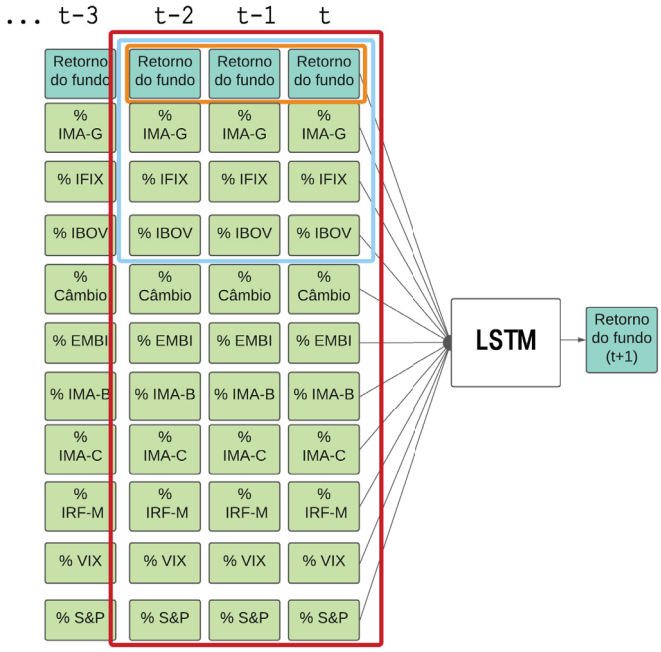


Figura 13. Seleção das entradas dos modelos para janela temporal de 3 dias. Fonte: Elaborado pelo autor.

Os dados normalmente estão dispostos em 2 dimensões: amostras e variáveis (atributos). A entrada de dados em uma rede LSTM tem uma particularidade, ela tem 3 dimensões. Além das amostras e variáveis, existe uma dimensão de janela temporal. Nesse caso, os dados passam por uma remodelagem conforme o tamanho da janela temporal escolhida.

A variação do tamanho da janela temporal faz com que a quantidade de dados disponível em cada período varie em alguns poucos dias. Por esse motivo, e para garantir que mesmo nas diferentes configurações as redes iriam executar os experimentos exatamente sobre os mesmos dados, foi necessário incluir um intervalo de 7 dias entre os dados de treino e validação, que pode ser observado na figura 12.

B. Treinamento dos modelos

Nessa etapa, foram utilizadas redes LSTM, escolhidas pela característica de preservar o gradiente e serem estudadas na literatura para fazer previsões em séries temporais [25].

1) Configuração:

Para todas as redes LSTMs utilizadas neste trabalho foi adotado a seguinte arquitetura: uma camada de entrada, uma camada oculta e a camada de saída. O tamanho do lote em 50 e o treinamento limitado a 500 épocas de treino, tendo como base trabalhos semelhantes que se limitaram seus treinamentos à valores semelhantes ou menores [6] [25].

A função de ativação escolhida para camada de saída foi a linear, pois, os valores de saída da regressão podem assumir valores negativos.

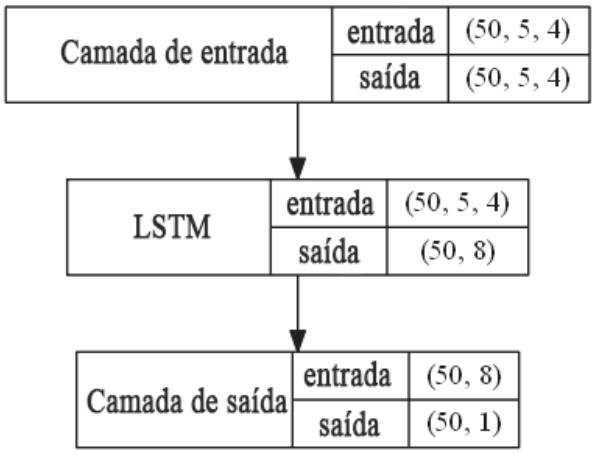


Figura 14. Exemplo da arquitetura da rede proposta, com uma janela temporal de 5 dias e usando o conjunto de dados 2, totalizando 435 parâmetros treináveis. Fonte: Elaborado pelo autor.

Este trabalho foi desenvolvido em Python (v3.7.10). A linguagem foi escolhida pela vasta gama de recursos e bibliotecas disponíveis para trabalhar com algoritmos de aprendizagem de máquina e rapidez no desenvolvimento.

A principal biblioteca usada foi o Keras (v2.3.1), uma conhecida API de aprendizado de máquina, que forneceu a implementação da LSTM. Pandas (v1.2.4) para manipulação e análise de dados. Statsmodels (v0.11.1) que forneceu a implementação do teste ADF e Tensorflow (v1.15.0) que foi necessário para implementação da função de custo customizada. Toda simulação executou em uma máquina local, fazendo uso da CPU (Intel Core i7-6700) com 32Gb de ram com sistema operacional Windows 10 PRO x64.

2) Hiperparâmetros:

Para um estudo mais abrangente, foram adotadas variações em alguns parâmetros de configuração. O número de neurônios na camada oculta variando entre 4, 8 e 16. O tamanho da janela temporal variando entre 1, 3, 5 e 7 dias.

Desta forma, a combinação das variações de configuração da rede, com os 3 conjuntos de dados vistos na sessão III-A4, resulta em 36 (3x4x3) modelos para cada um dos 33 fundos analisados.

3) Função de custo:

Como visto no capítulo II-D4 a função de custo MSE, por sua característica de penalizar valores discrepantes, é ideal para previsão de preços, onde *outliers* são esperados, porém, devem ser penalizados. Por outro lado, por ser uma função puramente estatística e simétrica, previsões negativas e positivas de mesma magnitude de erro tem o mesmo valor de custo [21]. Como neste trabalho, o sinal das previsões é o que vai guiar a tomada de decisão, ignorar essa informação seria uma falha de representação do modelo.

Em seu artigo, Lee (2007) exemplifica 4 funções para avaliar corretamente problemas de previsões de retornos financeiros. O problema é que a implementação delas para uso no treinamento da rede, esbarra na função de otimização vista no capítulo II-D5, que depende que estas funções sejam diferenciáveis e elas não são.

Uma função de custo alternativa foi proposta por Cheung (2020) e adaptada para o uso neste trabalho. Ela foi nomeada como *direction loss*, neste trabalho ela será referenciada como **DMSE**.

O conceito básico dela é penalizar a função MSE toda vez que a direção (sinal) do valor da previsão for diferente do esperado. A penalização é dada pelo coeficiente α que multiplica o valor da MSE original. Quanto maior o α maior é a penalização. Neste trabalho foi adotado $\alpha = 10$ o que significa que a cada erro de direção o custo é multiplicado por 10.

$$DMSE(\theta) = \begin{cases} MSE(\theta), & \text{se os sinais forem iguais} \\ MSE(\theta) * \alpha, & \text{caso contrário} \end{cases} \tag{20}$$

4) Parada antecipada:

Foi adotado um critério de parada antecipada (*early stopping*) para o treinamento dos modelos por dois motivos: Evitar o *overfitting* por excesso de treinamento e também para encurtar a etapa de treinamento das 1.188 (36 configurações x 33 fundos) redes.

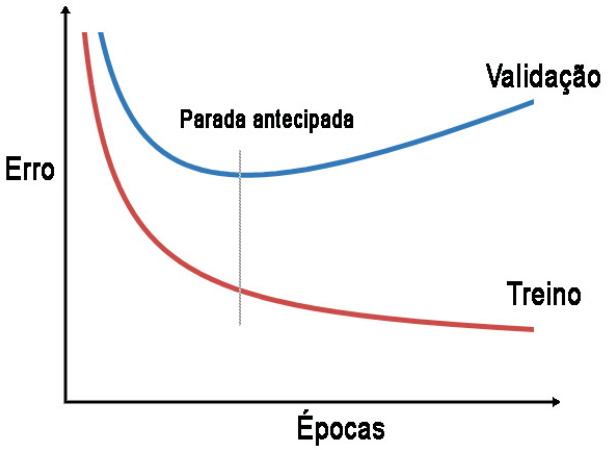


Figura 15. Exemplo do funcionamento da parada antecipada. Fonte: Elaborado pelo autor.

O critério escolhido foi parar o treinamento sempre que a taxa de erro no conjunto de validação não diminuir após 25 épocas. Quanto menor esse número maior o risco da rede encontrar-se em um platô no treinamento que eventualmente poderia melhorar depois de um tempo.

C. Estratégia de investimento

A estratégia de investimentos de um investidor hipotético é feita diariamente. Ao final de cada dia, no fechamento do mercado, a rede previamente treinada é alimentada com os dados atuais e faz a previsão do retorno de cada ativo para o próximo dia. Se a previsão de retorno da rede for maior que zero, o investidor assume uma posição comprada no ativo, esperando que o valor do ativo suba no dia seguinte. Se a previsão for menor que zero, o investidor assume uma posição vendida no papel, esperando que o valor do ativo caia. A predição do modelo pode ficar em zero, nesse caso, o investidor hipotético não faz nenhuma operação.
Se a previsão se mostrar correta, o lucro corresponderá a variação de preço efetiva do papel no dia, caso contrario, a variação do dia entrará como perda. Para simplificar a avaliação, custos de transação não são incluídos.

$$\text{retorno total} = \sum_{t=0}^{N} sign(f_t) * Y_{t+1} \tag{21}$$

Matematicamente o retorno da aplicação da estratégia é dada pela formula 21, onde $sign(f)$ é o sinal do valor de saída do modelo e Y o retorno efetivo do ativo.

Foi implementada também uma estratégia modificada, nela, a posição comprada só é assumida nos casos em que a previsão de retorno for $> 0,1\%$ e a posição vendida quando o retorno for $< -0,1\%$. Os casos em que a previsão ficar no intervalo de [-0,1%, 0,1%] representaria uma indefinição da previsão, uma vez que os valores estão próximos à zero, então nenhum posicionamento é feito e é atribuído retorno zero para o dia. Com o retorno total sendo dado pela formula 22.

$$\text{retorno total}' = \sum_{t=0}^{N} \begin{cases} sign(f_t) * Y_{t+1} & \text{se} |f_t| > 0.1\% \\ 0 & \text{caso contrário} \end{cases} \tag{22}$$

Em outras palavras, em caso de dúvida da previsão do modelo, o investidor não faria nenhuma operação naquele dia.

1) Regressão ou classificação:

Um olhar mais atento poderia questionar se com a estratégia adotada o problema não passaria a ser de classificação em vez de uma regressão. Alguns trabalhos seguem a linha da classificação [12] com a saída do modelo sendo binária: 'compra' ou 'venda'.

Esse tratamento certamente simplificaria o problema, mas ao particionar uma saída continua como preços ou retornos para uma saída binária, existiria perda de informação e diminuição da variância dos dados.

Um investidor não deveria estar preocupado, ao menos a princípio, tão e somente em quantos dias ele acertou o movimento e sim qual foi o retorno financeiro que ele teve ao final de um período. Para tal, a amplitude do movimento é importante, mas a informação da amplitude seria perdida numa binarização da saída.

Por estes motivos, mesmo que a estratégia de investimentos adotada pareça assumir um formato binário, o problema de previsão de tendências de retorno ainda é melhor representado como um problema de regressão.

D. Métricas de avaliação

Para quantificar o desempenho dos modelos na regressão, são utilizados métricas que analisam as diferenças entre a série de valores previstos e a série de valores reais. Neste trabalho serão aplicados a já citada função de erro quadrático médio (MSE) e o coeficiente de determinação R^2 que é uma medida estatística que basicamente, indica quanto o modelo foi capaz de explicar os dados coletados, dado pela equação:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \tag{23}$$

onde N é número de observações, $\overline{y_i}$ média das observações e $\widehat{y_i}$ é o valor estimado de y_i.

Como já foi melhor detalhado no capítulo da função de custo, funções puramente estatísticas não são ideais no contexto deste problema, por isso, além das métricas MSE e R^2, foi utilizado a DMSE (fórmula 20) e outras 2 métricas citadas no artigo do Lee (2007) que não puderam ser implementadas como função de custo devido a não serem diferenciáveis, mas que funcionam para avaliação dos resultados.

São elas a *mean forecast trading returns* (MFTR) e *mean correct forecast directions* (MCFD):

$$MFTR = \frac{1}{N} * \sum_{t=0}^{N} sign(f_t) * Y_{t+1} \tag{24}$$

$$MCFD = \frac{1}{N} * \sum_{t=0}^{N} sign(f_t) * sign(Y_{t+1}) > 0 \tag{25}$$

A primeira, nada mais é do que o retorno médio diário visto no capítulo anterior. A segunda é a média da somatória do número de vezes que o sinal da previsão foi o mesmo do valor efetivo, ou simplesmente, a porcentagem de dias em que o modelo acertou o movimento correto dos ativos (alta ou baixa).

E. Base de comparação

As métricas avaliam e comparam os modelos entre si, mas é preciso ter uma base de comparação (*baseline*) para comparar o desempenho da estratégia de investimentos usando os modelos com o próprio mercado de ativos.

Por exemplo, um modelo que teve retorno médio de 0,1% ao dia por determinado período pode ser considerado bom? Depende, se o ativo analisado valorizou no mesmo período 0,2% ao dia, um investidor que tivesse comprado o ativo e

não feito nenhuma outra operação de compra e venda, teria tido o dobro de retorno.

É possível criar bases de comparação com diversas complexidades de estratégias, esse trabalho vai limitar-se a estratégias bem simples, a ideia aqui é apenas dar uma noção de desempenho geral do mercado no período analisado.

Para isso, foi definido 4 estratégias de comparação:

- ***Buy and Hold***: É um dos *benchmarks* mais usados na renda variável, consiste em ficar comprado no ativo durante todo o período, não realizando nenhuma outra operação;
- **Sempre vendido**: É a estratégia oposta ao *Buy and Hold*, onde o investidor acredita na queda constante e fica vendido no ativo durante todo o período, não realizando nenhuma outra operação;
- **Segue tendência**: Nessa estratégia o hipotético investidor acredita que a tendência atual vai continuar, ou seja, se hoje o ativo valorizou, amanhã deve continuar subindo;
- **Retorno a média**: Oposta da 'Segue tendência', onde o investidor acredita que se hoje o ativo valorizou, amanhã ele deverá cair.

Estas estratégias servirão como comparativo na avaliação da rentabilidade resultantes dos modelos treinados.

Tabela VII
BASE DE COMPARAÇÃO - LISTA DE FUNDOS

Fundo	B&H[1]	SV[2]	ST[3]	RM[4]
BBFI11B	-0,03%	0,03%	0,28%	-0,28%
BBPO11	-0,13%	0,13%	0,20%	-0,20%
BBRC11	-0,09%	0,09%	0,16%	-0,16%
BCFF11	-0,01%	0,01%	0,03%	-0,03%
BPFF11	-0,08%	0,08%	-0,02%	0,02%
BRCR11	-0,04%	0,04%	0,05%	-0,05%
CEOC11	-0,07%	0,07%	-0,14%	0,14%
CTXT11	-0,25%	0,25%	-0,03%	0,03%
FAMB11B	-0,47%	0,47%	0,15%	-0,15%
FCFL11	0,05%	-0,05%	-0,18%	0,18%
FIGS11	-0,04%	0,04%	-0,07%	0,07%
FPAB11	-0,08%	0,08%	-0,49%	0,49%
HGBS11	-0,02%	0,02%	0,07%	-0,07%
HGCR11	0,03%	-0,03%	0,01%	-0,01%
HGLG11	-0,03%	0,03%	0,10%	-0,10%
HGRE11	-0,03%	0,03%	-0,01%	0,01%
JSRE11	-0,02%	0,02%	-0,06%	0,06%
KNCR11	0,02%	-0,02%	-0,01%	0,01%
MBRF11	-0,19%	0,19%	-0,17%	0,17%
MXRF11	0,00%	0,00%	0,00%	0,00%
ONEF11	0,01%	-0,01%	-0,18%	0,18%
PQDP11	-0,01%	0,01%	-0,21%	0,21%
RBBV11	0,06%	-0,06%	-0,15%	0,15%
RBED11	-0,04%	0,04%	0,00%	0,00%
RBVA11	-0,10%	0,10%	0,02%	-0,02%
RCRB11	-0,04%	0,04%	-0,06%	0,06%
RNGO11	-0,07%	0,07%	-0,09%	0,09%
SDIL11	0,05%	-0,05%	-0,05%	0,05%
SHPH11	0,03%	-0,03%	-0,27%	0,27%
THRA11	0,04%	-0,04%	-0,07%	0,07%
VLOL11	0,02%	-0,02%	-0,09%	0,09%
XPCM11	-0,11%	0,11%	0,53%	-0,53%
XTED11	-0,01%	0,01%	-0,22%	0,22%

[1]Buy and Hold, [2]Sempre vendido, [3]Segue tendência, [4]Retorno a média

IV. RESULTADOS E DISCUSSÕES

Nessa sessão serão descritos e discutidos os resultados obtidos com o uso das metodologias descritas no capítulo anterior. Para uma visualização mais assertiva, dado que algumas tabelas têm muitos valores, os mais relevantes foram destacados em negrito e os piores valores coloridos em vermelho.

O primeiro passo foi estabelecer as bases de comparação descritas no capítulo III-E. A tabela VII contem lista cada um dos 33 fundos selecionados para este estudo com seus respectivos resultados no período do conjunto de teste.

Para tentar mitigar que o acaso tenha grande interferência nos resultados, uma vez que os dados utilizados neste trabalho estão sujeitos a muito ruído, os resultados a seguir serão apresentados usando a média da regressão dos modelos de cada um dos 33 fundos.

Tabela VIII
BASE DE COMPARAÇÃO - MÉDIA DOS 33 FUNDOS

	B&H	SV	ST	RM
Acertos	-	-	44,45%	50,83%
Retorno médio	-0,050%	0,050%	-0,029%	0,029%

Nas tabelas IV, V e VI são apresentados os resultados médios das redes com as diferentes configurações de hiperparâmetros, para os 3 conjuntos de entrada.

Na tabela IV mostra os resultados dos modelos que receberam apenas o próprio retorno do fundo como entrada. Em média os resultados obtidos usando uma janela temporal de 5 e 7 dias se mostraram melhores.

Os resultados da tabela V são com os modelos que além do próprio fundo, receberam 3 indicadores como entrada. O retorno médio e porcentagem de acertos foram levemente superiores aos da tabela anterior, mas com uma piora significativa da métrica R^2.

Na tabela VI é apresentado os resultados do uso de todos os indicadores como entrada. Ela demonstra uma piora média de todos os indicadores utilizados, apresentando até um caso de retorno médio negativo.

Tabela IV
CONJUNTO DE ENTRADA 1 - MÉDIA DOS 33 FUNDOS

camada oculta	4				8				16			
janela temporal	1	3	5	7	1	3	5	7	1	3	5	7
R^2	-1,54%	**-2,98%**	-1,23%	-1,92%	-1,47%	-1,92%	-1,11%	-1,41%	-1,94%	-1,30%	**-1,61%**	**-1,84%**
MSE x10^4	1,893	**1,906**	1,840	1,850	1,844	1,843	1,832	1,837	1,883	1,840	**1,840**	**1,838**
DMSE x10^3	1,256	**1,289**	1,090	1,107	1,052	1,034	1,046	1,053	1,279	1,081	**1,031**	**1,022**
Acertos	49,39%	**48,58%**	49,55%	48,94%	49,62%	49,11%	48,76%	49,18%	49,27%	49,12%	**49,11%**	**48,85%**
Retorno médio	0,026%	**0,002%**	0,049%	0,035%	0,052%	0,052%	0,038%	0,044%	0,010%	0,040%	**0,058%**	**0,053%**

Tabela V
CONJUNTO DE ENTRADA 2 - MÉDIA DOS 33 FUNDOS

camada oculta	4				8				16			
janela temporal	1	3	5	7	1	3	5	7	1	3	5	7
R^2	-3,45%	**-3,73%**	**-2,03%**	-3,36%	**-1,60%**	-2,58%	**-2,85%**	-1,99%	-1,60%	-2,00%	-2,65%	-2,92%
MSE x10^4	1,888	**1,916**	**1,841**	1,897	**1,842**	1,864	**1,843**	1,856	1,849	1,847	1,874	1,865
DMSE x10^3	1,177	**1,261**	**1,081**	1,221	**1,029**	1,089	**1,053**	1,168	1,110	1,094	1,198	1,161
Acertos	49,73%	**49,44%**	**49,48%**	48,76%	**49,89%**	49,32%	**48,76%**	49,15%	50,03%	48,77%	49,59%	48,67%
Retorno médio	0,049%	**0,022%**	**0,053%**	0,032%	**0,064%**	0,045%	**0,047%**	0,045%	0,044%	0,034%	0,041%	0,043%

Tabela VI
CONJUNTO DE ENTRADA 3 - MÉDIA DOS 33 FUNDOS

camada oculta	4				8				16			
janela temporal	1	3	5	7	1	3	5	7	1	3	5	7
R^2	-5,75%	-7,47%	**-13,24%**	-5,09%	-6,56%	-5,92%	-4,69%	-5,79%	**-2,53%**	**-2,80%**	-3,77%	-5,48%
MSE x10^4	1,903	1,915	**2,015**	1,895	1,925	1,912	1,891	1,917	**1,867**	**1,874**	1,888	1,908
DMSE x10^3	1,203	1,174	**1,405**	1,119	1,290	1,106	1,074	1,267	**1,080**	**1,033**	1,194	1,237
Acertos	48,82%	48,30%	**48,14%**	47,45%	48,70%	48,61%	47,98%	47,64%	**49,36%**	**48,32%**	47,82%	47,58%
Retorno médio	0,015%	0,022%	**-0,003%**	0,019%	0,005%	0,034%	0,039%	0,011%	**0,049%**	**0,040%**	0,011%	0,005%

No geral, é possível fazer algumas constatações iniciais. Mesmo na melhor configuração o retorno médio de 0,064% ao dia não foi muito superior a melhor *baseline* que foi de 0,050%. Os piores resultados se concentraram no uso de 4 nós na camada oculta e que no geral o conjunto de entrada 3 foi consideravelmente pior que os outros dois.

Outro fato que chamou atenção foi a métrica R^2 ficou negativa para todos as configurações, em média. Informação que poderia indicar que os modelos tiveram algum tipo de problema durante a fase de treinamento. A confirmação dessa suspeita aconteceu quando os modelos foram apresentados aos dados de treinamento e o R^2 continuou negativo (na média dos 33 fundos) para todas as configurações de hiperparâmetros e conjuntos de entrada, um sinal claro de *underfitting*. Provavelmente os modelos não foram capazes de encontrar relações entre as entradas fornecidas e a saída esperada.

Uma nova filtragem dos fundos foi realizada, selecionando apenas aqueles que durante a fase de treinamento apresentaram R^2 acima de 1%, na tentativa de selecionar apenas os fundos nos quais a entrada tinha alguma significância estatística para o objetivo. Essa seleção resultou em 6 fundos: BPFF11, FPAB11, KNCR11, MBRF11, THRA11 e XPCM11.

Os resultados a seguir serão apresentados usando a média apenas desses 6 fundos selecionados, excluindo as colunas referentes a camada oculta com 4 nós e o conjunto de entrada 3 que tiveram os piores resultados no experimento anterior. A base de comparação foi novamente calculada, sem apresentar grandes diferenças com a anterior.

Tabela IX
BASE DE COMPARAÇÃO - MÉDIA DOS 6 FUNDOS COM $R^2 > 1\%$ NA FASE DE TREINAMENTO

	B&H	SV	ST	RM
Acertos	-	-	43,33%	49,42%
Retorno médio	-0,067%	0,067%	-0,037%	0,037%

Olhando para os números, não há grande diferença em relação a médias dos 33 fundos iniciais da tabela VIII

Tabela X
CONJUNTO DE ENTRADA 1 - MÉDIA DOS 6 FUNDOS COM $R^2 > 1\%$ NA FASE DE TREINAMENTO

camada oculta	8				16			
janela temporal	1	3	5	7	1	3	5	7
R^2	0,19%	**0,92%**	0,58%	1,09%	**0,26%**	1,27%	0,76%	**1,34%**
MSE x10^4	2,451	**2,424**	2,412	2,408	**2,441**	2,406	2,411	**2,395**
DMSE x10^3	1,562	**1,346**	1,364	1,452	**1,568**	1,369	1,364	**1,359**
Acertos	50,92%	**50,58%**	49,83%	50,42%	**50,83%**	50,33%	50,08%	**51,00%**
Retorno médio	0,109%	**0,168%**	0,151%	0,157%	**0,109%**	0,159%	0,165%	**0,181%**

Tabela XI
CONJUNTO DE ENTRADA 2 - MÉDIA DOS 6 FUNDOS COM $R^2 > 1\%$ NA FASE DE TREINAMENTO

camada oculta	8				16			
janela temporal	1	3	5	7	1	3	5	7
R^2	-0,81%	-0,74%	**-1,55%**	-2,54%	-0,75%	**-0,42%**	-1,78%	**-2,55%**
MSE x10^4	2,455	2,443	**2,422**	2,438	2,448	**2,441**	2,433	**2,431**
DMSE x10^3	1,502	1,374	**1,564**	1,542	1,504	**1,364**	1,476	**1,363**
Acertos	52,08%	49,42%	**47,58%**	48,00%	52,00%	**49,58%**	47,75%	**49,00%**
Retorno médio	0,150%	0,156%	**0,094%**	0,113%	0,152%	**0,159%**	0,115%	**0,166%**

A tabela X mostra um resultado bastante superior em comparação ao anterior contendo todos os fundos. Melhorando

em todas as métricas, R^2 positivo para todas as configurações, porcentagem de acertos acima de 50% na maioria dos casos e com a melhor configuração resultando em um retorno médio de **0,181%** ao dia, quase três vezes maior que a melhor *baseline* que foi o 'Sempre vendido' com **0,067%**.

A tabela XI também mostra uma melhora significativa de todas as métricas quando comparados aos da tabela V. Porém, com resultados piores que os que foram treinados usando apenas o conjunto de entrada 1.

Os resultados das tabelas X e XI deixam claros que a metodologia inicialmente usada não funciona para todos os fundos porque que muitos deles sofreram com *underfitting*. Portanto, uma seleção mais criteriosa se faz necessária para seleção de fundos cujos modelos sejam mais significativos.

Todos os resultados apresentados até aqui, seguiram a primeira estratégia descrita na sessão III-C, na qual, caso a previsão de retorno da rede for maior que zero, o investidor assume uma posição comprada no ativo, caso contrário assume uma posição vendida. Como consequência, o hipotético investidor teria que tomar uma posição todos os dias.

Mas o que aconteceria com os resultados, caso ele tivesse a opção de não operar caso o sinal retornado do modelo não fosse significativo o suficiente? Para responder a esta questão, foi adotado a estratégia modificada também detalhada na sessão III-C. Com ela, nos dias que a previsão ficar no intervalo de [-0,1%, 0,1%], que de certa forma representa uma incerteza do modelo, o investidor não faria nenhuma operação.

Tabela XII
CONJUNTO DE ENTRADA 1 - MÉDIA DOS 6 FUNDOS - ESTRATÉGIA MODIFICADA

camada oculta	8				16			
janela temporal	1	3	5	7	1	3	5	7
Dias operando	62,17%	**66,83%**	72,50%	70,42%	62,67%	67,17%	**72,50%**	**71,25%**
Acertos	54,34%	**53,17%**	51,67%	51,59%	54,03%	52,40%	**51,46%**	**52,33%**
Retorno médio	0,155%	**0,187%**	0,166%	0,154%	0,150%	0,179%	**0,184%**	**0,177%**

Uma comparação com a tabela X mostra uma melhora na porcentagem de acertos e retornos médios em quase todas as configurações.

Tabela XIII
CONJUNTO DE ENTRADA 2 - MÉDIA DOS 6 FUNDOS - ESTRATÉGIA MODIFICADA

camada oculta	8				16			
janela temporal	1	3	5	7	1	3	5	7
Dias operando	64,58%	**70,83%**	75,75%	78,25%	65,25%	70,50%	75,50%	77,58%
Acertos	51,69%	**51,05%**	48,09%	50,18%	52,23%	50,99%	47,77%	49,44%
Retorno médio	0,144%	**0,163%**	0,126%	0,148%	0,143%	0,148%	0,111%	0,124%

Uma comparação com a tabela XI mostra uma leve melhora na porcentagem de acertos.

Os resultados dessa modificação podem ser vistos nas tabelas XII, XIII e se mostraram promissores. Principalmente com o conjunto de entrada 1, que melhorou tanto os índices de acerto como retorno médio das operações. No conjunto 2 foi observado apenas uma leve melhora porcentagem de acertos com o retorno médio praticamente constante. Outra vantagem que não foi considerada nesse trabalho mas que na prática faz diferença, é a redução nos custos operacionais, uma vez que reduziu para aproximadamente 70% os dias de operação, sem afetar o retorno obtido. O resultado da melhor configuração obteve um retorno médio de **0,187%** ao dia, que durante os 200 dias de testes resultaria em um retorno total de **37,41%** no período. Bastante superior aos *baseline* propostos. O índice de acerto ficou pouco acima dos **52%** em média para conjunto de entrada 1, o que é compreensível diante do problema analisado, em que cada ponto percentual pode se traduzir em elevados ganhos financeiros.

Um último experimento foi feito, com a inclusão do período inicialmente excluído devido à pandemia do COVID-19 nos dados de teste. São aproximadamente 100 dias que apresentaram uma volatilidade dos retornos diários acima da média histórica. Como esses dias não tinham sido incluídos nos testes anteriores, as bases de comparação foram novamente calculadas e apresentadas na tabela XIV.

Tabela XIV
BASE DE COMPARAÇÃO - MÉDIA DOS 6 FUNDOS COM $R^2 > 1\%$ NA FASE DE TREINAMENTO - PERÍODO DA PANDEMIA INCLUÍDO

	B&H	SV	ST	RM
Acertos	-	-	43,78%	49,56%
Retorno médio	-0,032%	0,032%	-0,005%	0,005%

Retorno médio das bases de comparação foram bem mais modestos quando o período da pandemia foi incluído.

Tabela XV
CONJUNTO DE ENTRADA 1 - MÉDIA DOS 6 FUNDOS - PERÍODO DA PANDEMIA INCLUÍDO

camada oculta	8				16			
janela temporal	1	3	5	7	1	3	5	7
R^2	-1,35%	-2,26%	**-3,29%**	-3,62%	**-1,37%**	-1,98%	-3,27%	**-3,47%**
MSE x10^4	4,350	4,364	**4,388**	4,406	**4,350**	4,357	4,399	**4,411**
DMSE x10^3	2,904	2,957	**3,027**	3,067	**2,920**	2,917	3,035	**3,032**
Acertos	51,00%	50,17%	**49,94%**	50,78%	**50,89%**	49,89%	50,06%	**50,78%**
Retorno médio	0,103%	0,082%	**0,074%**	0,096%	**0,104%**	0,090%	0,082%	**0,104%**

Tabela XVI
CONJUNTO DE ENTRADA 2 - MÉDIA DOS 6 FUNDOS - PERÍODO DA PANDEMIA INCLUÍDO

camada oculta	8				16			
janela temporal	1	3	5	7	1	3	5	7
R^2	0,46%	1,90%	2,21%	1,29%	**0,61%**	2,04%	**2,93%**	**1,34%**
MSE x10^4	4,302	4,251	4,221	4,274	**4,288**	4,251	**4,189**	**4,258**
DMSE x10^3	2,741	2,484	2,435	2,565	**2,743**	2,502	**2,335**	**2,457**
Acertos	53,11%	51,72%	51,39%	51,50%	**52,89%**	51,83%	**51,56%**	**51,72%**
Retorno médio	0,184%	0,226%	0,233%	0,235%	**0,181%**	0,231%	**0,254%**	**0,259%**

Os resultados das tabelas XV e XVI devem ser comparados aos das tabelas X e XI respectivamente. Os resultados médios dos modelos usando o conjunto de entrada 1 com o período da pandemia incluído, tiveram um retorno médio pior, apesar das porcentagens de acertos tenham se mantidos quase nos mesmos patamares. Entretanto, os resultados mais interessantes são apresentados na tabela XVI, no qual, além de serem superiores aos seus pares sem o período pandêmico, pela primeira vez, foram observados vantagens no uso do conjunto de entrada 2 em relação ao conjunto 1. O uso de indicadores macroeconômicos ajudaram os modelos a performarem melhor nesse cenário de maiores incertezas e volatilidade de tal forma que a melhor configuração obteve um retorno médio de **0,259%** ao dia durante os 300 dias de testes, o que resultaria

em um retorno total de **77,7%** no período, mais de oito vezes superior a melhor base de comparação do período.

Por fim, uma decisão que se mostrou acertada foi o uso da função de custo DMSE para otimização dos modelos. Era esperado que um menor erro, resultasse em um maior retorno médio, portanto, uma correlação negativa. Fato esse que pode ser visto na tabela XVII, onde o DMSE obteve a melhor correlação negativa dentre as demonstradas nesse trabalho.

Tabela XVII
CORRELAÇÃO ENTRE AS FUNÇÕES DE CUSTO E O RETORNO MÉDIO

Função de custo	MSE	MAE	DMSE
Correlação	-0,239	0,096	-0,276

Na tabela XVIII foram selecionados resultados que demonstram essa melhor correlação na pratica.

Tabela XVIII
SELEÇÃO PARA EXEMPLIFICAR A CORRELAÇÃO ENTRE O RETORNO MÉDIO E AS MÉTRICAS USADAS

R^2	-1,61%	-1,30%	-1,60%	-2,85%
MSE x10^4	**1,840**	**1,840**	**1,842**	**1,843**
DMSE x10^3	**1,031**	**1,081**	**1,029**	**1,053**
Acertos	49,11%	49,12%	49,89%	48,76%
Retorno médio	**0,058%**	**0,040%**	**0,064%**	**0,047%**

Foram escolhidos propositalmente resultados com praticamente o mesmo MSE para facilitar a comparação.

V. CONSIDERAÇÕES FINAIS

Este trabalho teve o objetivo de propor um modelo de previsão do mercado de fundos de investimentos imobiliários através de uma rede LSTM, alimentada com os dados históricos do próprio fundo e também verificar se o uso de indicadores macroeconômicos auxiliariam a capacidade preditiva dos modelos.

Com os resultados das previsões obtidas no capítulo anterior, foi possível verificar que as diversas configurações têm diferenças significativas na capacidade de previsão. Outra constatação foi de que o modelo proposto não funcionou para todos os fundos inicialmente selecionados, pois muitos deles sofreram com *underfitting*. Mas que após uma nova seleção dos fundos, os resultados se mostraram promissores, tendo um retorno médio diário acima das bases de comparação estabelecidas, com as melhores configurações obtendo um retorno de três a oito vezes maior para o período em relação às bases de comparação propostas. Principalmente com o uso da estratégia mais conservadora, na qual, nos dias em que a previsão apresente incerteza, não é feita nenhuma operação, com algumas configurações alcançando um índice de acerto do sinal das operações maior que 53%.

A questão do uso de indicadores macroeconômicos, parece não ter tido uma resposta definitiva. Para quase todos os experimentos realizados, o subconjunto de indicadores usados neste trabalho pareciam indicar que não eram úteis na previsão. Com os modelos que utilizaram o conjunto de entrada 1 (que tinha apenas o histórico de retornos do próprio fundo como entrada), performando melhor que os demais. A exceção aconteceu quando nos dados de teste, foram incluídos o período da pandemia do COVID-19, nesse caso o conjunto de entrada 2 teve um desempenho muito melhor do que tinha sido observado até então.

O motivo dessa melhora tão significativa merece ser analisada em trabalhos futuros, foram levantadas duas hipóteses que podem ajudar a explicá-la: Num período com muitas incertezas, indicadores macroeconômicos passam a ter relevância maior para predição, pois os investidores estariam mais preocupados com o cenário macroeconômico do que com o desempenho individual dos fundos; A lacuna deixada ao excluir o período atrapalhou os demais experimentos.

Como contribuição, o uso da função de custo DMSE mostrou que pode ser útil, obtendo a melhor correlação negativa com o objetivo principal que é maximizar o lucro. Apesar de ter sido a melhor função, o valor de -0,28 indica uma correção baixa, o que deixa margem para novos trabalhos explorarem o assunto, uma vez essa função tem um parâmetro ajustável α. Se α for setado com um valor baixo, a amplitude do erro tem maior importância. Caso α seja setado com um valor alto, a direção do erro passa a ter mais relevância. Encontrar o ajuste ideal não é trivial, uma vez que o investidor para ter lucro precisa acertar a direção do movimento e o lucro vem da amplitude desse movimento.

Outro ponto a ser observado e trabalhado no futuro é a definição de critérios de seleção e tratamento mais individualizado para cada fundo. Como visto, dos 33 fundos inicialmente selecionados, apenas 6 tiveram um treinamento efetivo. A abordagem usada neste trabalho, de tratar os fundos de forma homogênea, usando para todos os mesmos indicadores macroeconômicos provavelmente não é adequada. Uma busca de características mais significativas também se faz necessária. Uma etapa inicial de clusterização pode ajudar na obtenção de resultados mais consistentes para uma gama maior de fundos.

Uma limitação do estudo foi a não inclusão dos custos operacionais na contabilização dos resultados finais. Parte da motivação dessa decisão foi devido ao fato das principais corretoras do país, não cobrarem corretagem em operações com FIIs [17]. Não tendo portando, custos fixos nas operações, com as únicas taxas incidindo sobre o volume financeiro movimentado, que devem ser considerados em algum momento, mas é algo muito particular de cada investidor.

Os resultados encontrados contribuem para o tema da criação de modelos preditivos com o intuito de ajudar operações com Fundos Imobiliários. Esses resultados acima da média, caso confirmados em um experimento prático, vão no sentido de contrapor a hipótese do mercado eficiente até mesmo em sua forma fraca.

Por fim ressalta-se que a área de previsões de mercados financeiros é muito visada e estudada, o que provê uma infinidade de outros caminhos a serem tomados a partir este trabalho.

Referências

[1] D. S. Academy. Deep learning book, 2021. URL http://www.deeplearningbook.com.br.

[2] B3. Uma análise da evolução dos investidores pessoas físicas na b3, 2020. URL http://www.bmf.com.br/portal/pages/newsletter/BMFBOVESPA/Downloads/Estudo_PF-final.pdf.

[3] B3. Boletim mensal fundos imobiliários, 2020. URL http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/fundos-de-investimentos/fii/boletim-mensal/.

[4] Banco central. Taxas de juros básicas – histórico, 2020. URL https://www.bcb.gov.br/controleinflacao/historicotaxasjuros.

[5] A. F. Bueno. Análise empírica do 'dividend yield' das ações brasileiras. *FEA/USP*, 2000. URL https://teses.usp.br/teses/disponiveis/12/12136/tde-01032002-120517/pt-br.php. Dissertação de Mestrado.

[6] A. C. M. P. Caio Mário Mesquita, Renato Oliveira. Utilização de uma rede neural lstm e testes da razão da variância para previsões em séries de ativos da bovespa. *Universidade Federal de Minas Gerais*, 2019. URL http://www.comp.ita.br/labsca/waiaf/papers/CaioMesquita_paper_11.pdf.

[7] L. G. Canto. Análise de notícias do mercado financeiro utilizando processamento de linguagem natural e aprendizado de máquina para decisões de swing trade. *Universidade Federal do Rio de Janeiro*, 2020. URL http://monografias.poli.ufrj.br/monografias/monopoli10031581.pdf.

[8] Z. Cheung. Customize loss function to make lstm model more applicable in stock price prediction, 2020. URL https://towardsdatascience.com/customize-loss-function-to-make-lstm-model-more-applicable-in-stock-price-prediction-b1c50e50b16c.

[9] CMV. Fundo de investimento imobiliário - caderno nº 6, 2010. URL https://www.investidor.gov.br/portaldoinvestidor/export/sites/portaldoinvestidor/publicacao/Cadernos/CVM-Caderno-6.pdf.

[10] R. A. da Silva. Estimação dinâmica do beta do modelo capm em fundos de ações: uma aplicação do filtro de kalman. *Universidade de São Paulo*, 2007. URL https://repositorio.usp.br/result.php?filter[]=unidadeUSP:%22MODMATFIN%22. Dissertação de Mestrado.

[11] R. de Losso Da Silveira Bueno. *Econometria De Séries Temporais*. Cengage, São Paulo, 2008. ISBN 8-522-10642-8.

[12] F. dos Santos Giacomel. Um método algorítmico para operações na bolsa de valores baseado em ensembles de redes neurais para modelar e prever os movimentos dos mercados de ações. *Universidade Federal do Rio Grande do Sul*, 2016. URL http://hdl.handle.net/10183/134586. Dissertação de Mestrado.

[13] R. E. P. P. Fabrício Soares, Rejane Frozza. Predição de séries financeiras utilizando wavelets e redes neurais: ummodelo para os fundos de investimentos imobiliários. *Universidade do Estado do Rio de Janeiro - UERJ*, pages 36–53, 2008. URL https://www.e-publicacoes.uerj.br/index.php/cadest/article/view/15732. Cadernos do ime.

[14] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, pages 383–417, 1970.

[15] M. A. F. Finocchio. *Noções de redes neurais artificiais*. eBook, 2014. URL http://paginapessoal.utfpr.edu.br/mafinocchio/labsi-laboratorio-de-seguranca-e-iluminacao/redes-neurais-artificiais/.

[16] G1. Poupança perde para a inflação em 2020 e tem pior rentabilidade em 18 anos, 2020. URL https://g1.globo.com/economia/noticia/2021/01/12/poupanca-perde-para-a-inflacao-em-2020-tem-pior-rentabilidade-em-18-anos.ghtml.

[17] Globo. Compare as principais taxas cobradas pelas corretoras e plataformas de investimentos, 2020. URL https://valorinveste.globo.com/produtos/servicos-financeiros/noticia/2020/10/06/compare-as-principais-taxas-cobradas-pelas-corretoras-e-plataformas-de-investimentos.ghtml.

[18] A. Géron. *Mãos à obra: aprendizado de máquina com Scikit-Learn TensorFlow*. Alta Books; 1ª edição, 2019.

[19] J. Heaton. *Introduction to the Math of Neural Networks*. Heaton Research, eBook, 2012.

[20] R. J. Hyndman and G. Athanasopoulos. Forecasting: principles and practice, 2018. URL https://otexts.com/fpp2/.

[21] T.-H. Lee. Loss functions in time series forecasting. *University of California*, 2007. URL http://www.faculty.ucr.edu/~taelee/paper/lossfunctions.pdf. Artigo.

[22] K. C. d. S. R. Luciano Ferreira Carvalho, Osvaldo Silvio Giachero. Uma crítica a hipótese da eficiência de mercado: Análise dos casos ambev e submarino. *Revista de Gestão, Finanças e Contabilidade*, 2013. URL https://revistas.uneb.br/index.php/financ/article/view/95.

[23] A. luiz de paula carvalho. Modelagem econométrica para forecasting do ifix. *FGV EPGE*, 2019. URL https://bibliotecadigital.fgv.br/dspace/handle/10438/28498. Dissertação de Mestrado.

[24] B. Malkiel. *A Random Walk Down Wall Street*. W. W. Norton and Company, Inc., EUA, 1973. ISBN 0-393-06245-7.

[25] D. K. F. Matsumoto. Estudo em séries temporais financeiras utilizando redes neurais recorrentes. *Universidade Federal de Alagoas*, 2020. URL http://www.repositorio.ufal.br/handle/riufal/6813. Dissertação (Mestrado em Modelagem Computacional do Conhecimento) – Instituto de Computação, Programa de Pós Graduação em Modelagem do Conhecimento.

[26] J. Mendel and R. McLaren. 8 reinforcement-learning control and pattern recognition systems. In J. Mendel and K. Fu, editors, *Adaptive, Learning and Pat-*

tern Recognition Systems, volume 66 of *Mathematics in Science and Engineering*, pages 287–318. Elsevier, 1970. doi: https://doi.org/10.1016/S0076-5392(08)60497-X. URL https://www.sciencedirect.com/science/article/pii/S007653920860497X.

[27] R. A. Mendes. *Fundos de Investimento Imobiliário: Aspectos Gerais e Princípios de Análise*. Novatec Editora, São Paulo, 2018. ISBN 8-575-22675-4.

[28] L. P. R. D. F. Parreiras. Arbitragem estatística e inteligência artificial. *Universidade de São Paulo*, 2007. URL https://www.ime.usp.br/~rvicente/LuizParreiras_MSc.pdf. Dissertação de Mestrado.

[29] R. A. D. Peter J. Brockwell. *Introduction to Time Series and Forecasting*. Springer, 3ª edição, 2016. ISBN 978-3319298528.

[30] R. D. Reed and R. J. Marks. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press, 1999.

[31] F. H. Rubim. Deep learning no mercado acionário brasileiro: fatores que possibilitam previsões consistentes para a tomada de decisão em condições de risco. *Universidade De São Paulo*, 2019. URL https://www.teses.usp.br/teses/disponiveis/12/12139/tde-17032020-183245/pt-br.php. Pós-Graduação.

[32] F. Soares. Modelo de predição financeira utilizando wavelets e redes neurais artificiais. *Universidade de Santa Cruz do Sul*, 2009. URL http://hdl.handle.net/11624/341. Dissertação de Mestrado.

[33] K. Y. Yokoyama. Os fundos de investimento imobiliário listados em bolsa e as variáveis que influenciam seus retornos. *Universidade Federal do Espírito Santo*, 2014. URL https://repositorio.ufes.br/handle/10/1307. Dissertação de Mestrado.