Dissertação apresentada à Pró-Reitoria de Pós-Graduação do Instituto Tecnológico de Aeronáutica, como parte dos requisitos para obtenção do título de Mestre em Ciências no Programa de Pós-Graduação em Engenharia Eletrônica e Computação, Área de Informática.

Luiz Alfredo Zenon da Mata Caffé

DETECÇÃO DE FRAUDES EM CRIPTOMOEDAS UTILIZANDO MÉTODOS DE CLASSIFICAÇÃO DE SÉRIES TEMPORAIS BASEADOS EM REDES NEURAIS

Dissertação aprovada em sua versão final pelos abaixo assinados:

Prof. Dr. Cesar Augusto Cavalheiro Marcondes

Orientador

Prof. Dr. Pedro Teixeira Lacava

Pró-Reitor de Pós-Graduação

Campo Montenegro
São José dos Campos, SP - Brasil
2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Zenon da Mata Caffé, Luiz Alfredo
Detecção de Fraudes em criptomoedas utilizando métodos de classificação de séries temporais baseados em redes neurais / Luiz Alfredo Zenon da Mata Caffé.
São José dos Campos, 2021.
119f.

Dissertação de Mestrado – Curso de Engenharia Eletrônica e Computação. Área de Informática – Instituto Tecnológico de Aeronáutica, 2021. Orientador: Prof. Dr. Cesar Augusto Cavalheiro Marcondes.

1. Oferta Inicial de Moedas. 2. Criptomoedas. 3. Blockchain. 4. Séries Temporais. 5. Redes Neurais. I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

ZENON DA MATA CAFFÉ, Luiz Alfredo. **Detecção de Fraudes em criptomoedas utilizando métodos de classificação de séries temporais baseados em redes neurais**. 2021. 119f. Dissertação de Mestrado – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Luiz Alfredo Zenon da Mata Caffé
TITULO DO TRABALHO: Detecção de Fraudes em criptomoedas utilizando métodos de classificação de séries temporais baseados em redes neurais.
TIPO DO TRABALHO/ANO: Dissertação / 2021

Luiz Alfredo Zenon da Mata Caffé
Praça Marechal Eduardo Gomes, 50 Vila das Acácias,
12228-900 – São José dos Campos/SP - Brasil

DETECÇÃO DE FRAUDES EM CRIPTOMOEDAS UTILIZANDO MÉTODOS DE CLASSIFICAÇÃO DE SÉRIES TEMPORAIS BASEADOS EM REDES NEURAIS

Luiz Alfredo Zenon da Mata Caffé

Composição da Banca Examinadora:

Prof. Dr.	Adilson Marques da Cunha	Presidente	-	ITA
Prof. Dr.	Cesar Augusto Cavalheiro Marcondes	Orientador	-	ITA
Prof. Dr.	Paulo Marcelo Tasinaffo	Membro Interno	-	ITA
Prof. Dr.	Arlindo Flavio de Conceição	Membro Externo	-	ITA

ITA

"Todos morrem sozinhos. Mas se você significa algo para alguém, se você ajuda alguém ou ama alguém, se até mesmo uma única pessoa se lembra de você, então, talvez, você nunca morra de fato."

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, por ser o fôlego da vida.

Em segundo, aos meus pais, por terem me dado o fôlego da vida e me criado com valores que me permitiram chegar até aqui.

Ao Prof. Dr. Cesar Marcondes, meu Orientador, pela confiança depositada na realização deste trabalho.

E, por último, a todos os meus amigos que, de alguma forma, me auxiliaram para a consecução deste trabalho.

"Ideias, e somente ideias,
podem iluminar a escuridão."
— Ludwig Von Mises

Resumo

Este trabalho apresenta um método para a detecção de fraudes em criptomoedas, originadas a partir de uma Oferta Inicial de Moedas (*Initial Coin Offering* - ICO). Para isto, foram utilizados modelos preditivos, baseados em redes neurais, para a classificação de Séries Temporais, geradas a partir das tabelas de fluxo de transações na rede Ethereum. A primeira atividade de ICO foi executada em 2013 e alcançou o seu auge no primeiro semestre de 2018, com movimentações entre 7 e 12 bilhões de USD em todo o mundo. Todavia, estima-se que 78% das atividades de ICO são fraudulentas. Baseadas no comportamento de criptomoedas fraudulentas e não fraudulentas, bem como nas tabelas de transações das criptomoedas coletadas, foram desenvolvidas 5 séries temporais normalizadas, que deram entrada nos modelos de Redes Neurais Artificiais (RNA) dos tipos *Multi Layer Perceptron* (MLP), *Convolution Neural Network - Multi Layer Perceptron* (CNN-MLP) e *Long Short Term Memory - Multi Layer Perceptron* (LSTM-MLP) projetados para classificação. Ao final da pesquisa, foi obtido um valor de (*Recall*) de até 91% em alguns casos.

Abstract

This work presents a method of detecting fraud in cryptocurrencies, originated from an Initial Coin Offering (ICO). For this, predictive models were used, based on neural networks, for the classification of time series, generated from transaction flow tables in the Ethereum network. The first ICO activity was carried out in 2013 and reached its peak in the first half of 2018, with a turnover between 7 and 12 billion dollars worldwide. However, it is estimated that 78% of ICO activities are fraudulent. Based on the behavior of fraudulent and non-fraudulent cryptocurrencies, as well as on the transaction tables of the collected cryptocurrencies, 5 normalized time series were developed, which were input into the following types of Artificial Neural Networks (ANN): Multi Layer Perceptron (MLP); Convolution Neural Network - Multi Layer Perceptron (CNN-MLP); and Long Short Term Memory - Multi Layer Perceptron (LSTM-MLP) designed for classification. At the end of the research, a *Recall* value of up to 91% was obtained in some cases.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ICO	Oferta Inicial de Moedas (Initial Coin Offering)
P2P	Peer to Peer
RNA	Redes Neurais Artificiais
ReLU	Unidade Linear Retificada
MLP	Multilayer Perceptron
CNN	Redes Neural Convolucional (Convolution Neural Network)
LSTM	Long-Short Term Memory
tanh	Tangente Hiperbólica
ARIMA	Média Móvel Integrada Autoregressiva (Autoregressive Integrated Moving Average)

Lista de Símbolos

σ	Sigmóide
$\oplus$	Adição
$\otimes$	Multiplicação

Sumário

1 Introdução

Blockchain é uma das invenções mais inovadoras dos últimos anos e tem sido considerada como a tecnologia mais disruptiva na Internet, tendo sido concebida por uma pessoa anônima, com pseudônimo de Satoshi Nakamoto em 2008 (BECK; MÜLLER-BLOCH, 2017). Esta tecnologia consiste em uma arquitetura descentralizada, baseada em redes *peer-to-peer* e protocolos de consenso distribuído, que vêm solucionando vários desafios da comunidade de sistemas distribuídos, ao utilizar criptografia avançada para a construção de um livro-razão público e acessível por qualquer usuário (NAKAMOTO, 2019).

Dessa forma, nós de redes dividem responsabilidades pela validação de transações a serem inseridas, não havendo a necessidade de uma entidade central confiável para validá-las. Essa tecnologia se tornou o núcleo principal da criptomoeda Bitcoin, que teve grande sucesso nos últimos anos (ULRICH, 2017).

A tecnologia *Blockchain* possui diversas aplicabilidades, em áreas diversas de: Finanças, Verificação de Integridade, Governança, Internet das Coisas, Saúde, Educação, Privacidade e Segurança, Negócios e Indústria e Gerenciamento de Dados. Sua utilidade como banco de dados público e imutável traz muitas vantagens e características únicas. De fato, especificamente na área de Finanças (CASINO *et al.*, 2018), esta tecnologia possibilitou a criação de novas maneiras de financiamento coletivo e a criação das chamadas moedas digitais, cunhadas dentro da *Blockchain*.

Para esse tipo de aplicação de financiamento coletivo, além do livro-razão público, também foram necessárias outras inovações complementares, como os Contratos Inteligentes (*Smart Contracts*), os quais são programas de computador de propósito geral, hospedados na rede *blockchain* e executados de maneira independente e descentralizada, segundo uma lógica de intermediação (SZABO, 1997).

Além disso, esses contratos possuem o objetivo de orquestrar ações entre partes envolvidas, de modo a propiciar a execução de transações, mesmo que as partes não confiem umas nas outras (ABDELHAMID; HASSAN, 2019). A finalidade consiste em reduzir o uso de terceiros e reduzir custos em arbitração, bem como também traz consigo as outras propriedades de publicidade e imutabilidade. Isto possibilitou que a *blockchain* aumentasse o espectro de sua empregabilidade, servindo de plataforma para inúmeros novos tipos

de aplicações (CASINO *et al.*, 2018); entre elas, o foco deste trabalho é no financiamento coletivo para empresas que desejam captar recursos para iniciar seus projetos de negócio.

No mundo real, distante da nova realidade digital, as *startups* são empresas em fase inicial de formação de novos negócios. Elas representam novos empreendimentos que se propõem a apresentar soluções inovadoras, repetíveis e escaláveis para o mercado, necessitando levantar capital rapidamente para seus projetos. Embora exista um caminho normal para os novos empreendedores, de buscar um investidor anjo, ou um empréstimo bancário, e depois, investimentos mais substanciais em aceleradoras e venda de participação societária, mesmo assim a empresa não é oferecida ao grande público.

Como os investidores e bancos que financiam *startups* são em pequeno número, eles escolhem muito rigorosamente os empreendimentos, de forma que, mesmo que outros tenham alto potencial, podem não ser escolhidos. Sendo assim, o investimento em novas empresas não é algo amplamente disponível, nem democratizado. Somente as empresas que possuem seus negócios consolidados, geralmente após anos de investimentos iniciais, têm a possibilidade de serem oferecidas a uma bolsa de valores, em uma atividade chamada Oferta Pública Inicial (IPO), onde suas ações são compradas e vendidas por investidores e, assim, a empresa se capitaliza.

As IPOs no Brasil são administradas pela B3, que é a bolsa de valores oficial do Brasil. Nos anos com maior quantidade de IPOs, como em 2007, a Bolsa brasileira conseguiu o recorde de 64 IPOs, que somaram R$55,6 bilhões na época [1]. Ou seja, em qualquer ano normal, no máximo 50 empresas são abertas para o público no Brasil.

Além disso, a bolsa de valores brasileira é muito elitizada, com apenas 3 milhões de investidores no universo brasileiro [2], sendo 1/3 deles iniciando em 2019. Por outro lado, no mundo digital, casas de câmbio de Bitcoin, como a Binance[3], possuem mais de 15 milhões de usuários, lidando com um ativo ainda especulativo.

Uma das principais vantagens das criptomoedas é que elas são acessíveis para qualquer pessoa, independente do seu perfil de investimento, classe social e mesmo legislação do país. E sendo assim, com o advento dos contratos inteligentes, executados na *blockchain*, foi possível suprir a necessidade de popularizar e democratizar ainda mais o financiamento coletivo.

Portanto, por meio da atividade chamada de Oferta Inicial de Moedas (ICO), realizada através de *Smart Contracts*, torna-se possível realizar o equivalente à IPO. Os usuários de criptomoedas podem comprar o equivalente a ações (conhecidas por *tokens*) de um novo empreendimento, através de um intermediador automatizado (*smart contract*).

[1] https://conteudos.xpi.com.br/acoes/relatorios/onda-de-ipos-no-brasil-e-para-se-preocupar/
[2] http://www.b3.com.br/pt_br/noticias/investidores.htm
[3] https://www.binance.com/pt-BR/

A primeira atividade de ICO foi executada em 2013, quando a tecnologia de *smart contracts* estava no início. Após alguns anos, as atividades de ICO alcançaram o seu auge no primeiro semestre de 2018, impulsionadas pela alta das criptomoedas, obtendo movimentação entre 7 e 12 bilhões de dólares americanos em todo o mundo (CHOD; LYANDRES, 2019). Tal variação expressiva é explicada pelo fato destas atividades não serem regulamentadas na maioria dos países. Sendo assim, é difícil informar com precisão a quantidade de recursos efetivamente movimentada.

1.1 Motivação

Obviamente, tornar os investimentos mais populares e fáceis de realizar tem o seu lado bom, mas também, seu lado negativo. O público alvo se torna maior e mais inexperiente. Isto abre o campo para um aumento potencial de fraudes. Todo empreendedor, por mais simples que seja, pode colocar seu negócio em qualquer atividade de ICO e receber grande quantidade de recursos, caso ela seja bem sucedida. E o investidor inexperiente, por sua vez, pode colocar seus recursos em várias atividades de ICO, com pouco critério, imaginando que uma destas possa aumentar de valor em muitas vezes, proporcionalmente ao que foi inicialmente investido.

O primeiro caso significativo de fraude em criptomoedas foi o caso DAO[4]. A criptomoeda entrou no mercado em abril de 2016. A empresa que a desenvolveu teve como proposta permitir que as pessoas alugassem seus bens (carros, barcos, apartamentos), usando a tecnologia de *Blockchain*, sem intermediários. Porém, em junho do mesmo ano, o DAO sofreu ataque *hacker* e perdeu cerca de US$ 60 milhões em valor de mercado Ethereum. Outro levantamento realizado em 2018 indica que US$ 9,1 milhões foram perdidos por dia, em fraudes de criptomoedas.[5] Isto mostrou que o aspecto psicológico de corrida por oportunidade propiciada pelas criptomoedas ocasionou danos em muitos investidores.

Mais especificamente, nos sistemas de financiamento coletivo, não foi diferente. Em abril de 2018, um caso ganhou manchetes dos jornais como o maior escândalo de fraudes nessa área[6]. Acredita-se que duas ICOs, Pincoin e IFan, administradas por uma mesma empresa que opera no Vietnã, tenham furtado cerca de US$ 660 milhões de aproximadamente 32.000 investidores. O IFan prometia ser uma plataforma de mídia social para celebridades promoverem seus conteúdos para fãs, enquanto a Pincoin se propôs a entregar uma série de oito produtos, todos baseados em *Blockchain*.

Conforme o estudo publicado[7] pela Satis Group (empresa especializada em consultoria

[4]https://www.coindesk.com/understanding-dao-hack-journalists
[5]https://news.bitcoin.com/9-million-day-lost-cryptocurrency-scams/
[6]https://cointelegraph.com/news/unpacking-the-5-biggest-cryptocurrency-scams
[7]https://research.bloomberg.com/pub/res/d28giW28tf6G7T_Wr77aU0gDgFQ

financeira de criptomoedas), em julho de 2018, 81% de todas ICOs foram fraudulentas e somente 15% conseguiram ser listadas em uma casa de câmbio, a fim de serem negociadas no mercado. Portanto, o número relativo de fraudes foi excessivamente alto em 2018. Além disto, de acordo com a plataforma Google Trends[8], o auge da popularidade de buscas por fraudes em criptomoedas se deu no meio do ano de 2018, dois meses após a alta procura por ICOs, conforme Figura 1.1.

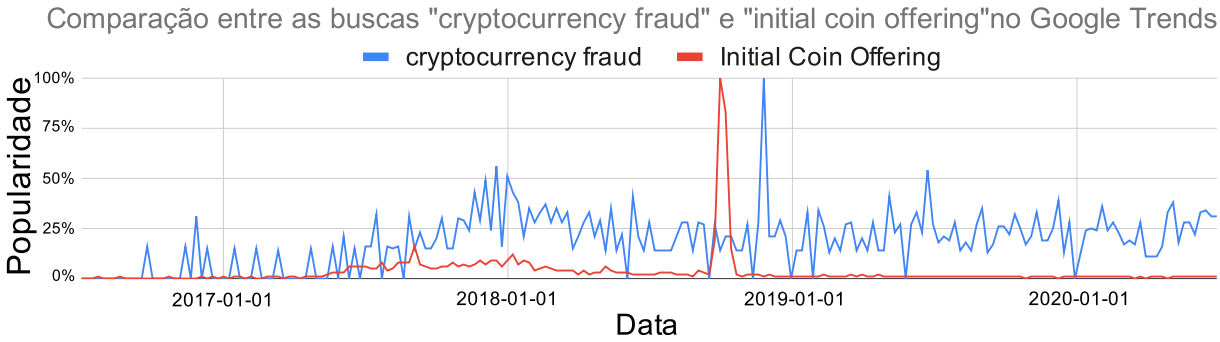


FIGURA 1.1 – Gráfico que indica o nível de popularidade das buscas. O indicador de 100% representa o pico da popularidade.

Portanto, apesar de a procura por este assunto ter diminuído nos últimos anos, continua representando algo pesquisado nos últimos meses. Portanto, encontrar uma solução que resolva ou diminua atividades fraudulentas constituiu-se a motivação para este estudo.

1.2 Objetivos e Contribuições

Este trabalho tem como objetivo apresentar um método para a detecção de fraudes em criptomoedas, originadas a partir de Oferta Inicial de Moedas (ICO). O Método será baseado na construção de diversos modelos preditivos, baseados em redes neurais, para a classificação de Séries Temporais, geradas a partir das tabelas de fluxo de transações ao longo da rede *Blockchain* Ethereum.

Adicionalmente, ao objetivo principal, foi realizado um levantamento bibliográfico sistemático entre junho de 2019 e junho de 2020 para contextualizar o problema, revelar as técnicas de estado da arte na detecção de fraudes em criptomoedas e embasar este trabalho.

As contribuições deste trabalho são: Análise Exploratória de Dados (EDA) dos bancos de dados das transações das criptomoedas coletadas, avaliação comparativa entre os desempenhos de modelos de classificação baseados em Redes Neurais Artificiais (RNAs) e aplicação de Séries Temporais para detectar fraudes nesta área.

[8]https://trends.google.com.br/

Entre os resultados obtidos, destaca-se a análise de 238 conjuntos de dados de ICOs de criptomoedas, sendo 136 fraudulentas e 102 não-fraudulentas. Foram utilizados 45 experimentos de classificação de Séries Temporais: 15 modelos de *Multilayer Perceptron* (MLP), 15 modelos de Redes Neurais Convolucionais (CNN-MLP) e 15 modelos de *Long-Short Term Memory* (LSTM-MLP). Todos estes apresentaram a capacidade de detectar fraudes em criptomoedas, chegando a alcançar um desempenho de 91% de *Recall*, para amostras de tempo de 20 dias após o lançamento da criptomoeda no mercado. Tal desempenho indica uma superioridade em comparação ao encontrado na literatura, (CHEN *et al.*, 2018) com 81% e (CHEN *et al.*, 2019) com 69%.

1.3 Organização da Dissertação

A Figura 1.2 apresenta o Método para a detecção de fraudes em criptomoedas, tema deste trabalho. Ela contém cinco passos sequenciais, os quais são embasados nos conhecimentos provenientes da Fundamentação Teórica e dos Trabalhos Relacionados.

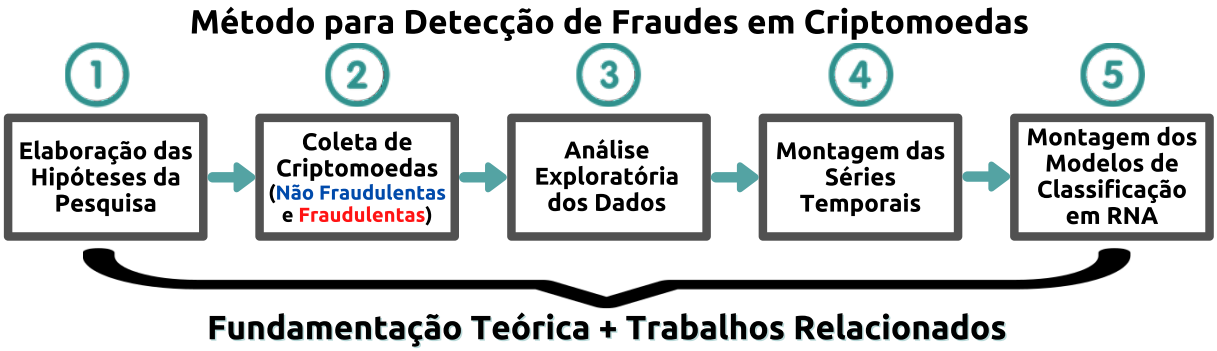


FIGURA 1.2 – Método para a Detecção de Fraudes em Criptomoedas.

Neste Capítulo 1, foi resumida uma introdução de contextualização da área e do problema a ser estudado, foram enunciados os objetivos e contribuições do trabalho e destacados os principais resultados a serem obtidos.

Em seguida, torna-se necessário, para a melhor compreensão deste trabalho, uma compreensão mínima sobre os assuntos: *Blockchain* Ethereum, Mercado de Criptomoedas, Séries Temporais e Redes Neurais. Estes assuntos são abordados no Capítulo 2 e compõem a Fundamentação Teórica.

A revisão sistemática da literatura e seus resultados, que estreitaram a relação com a proposta de solução do problema apresentado neste estudo, encontram-se detalhadas no Capítulo 3, compondo os Trabalhos Relacionados.

Em seguida, a metodologia desta pesquisa encontra-se no Capítulo 4, explicando os critérios de seleção das bases de dados, uma análise de caracterização dos dados e a

confecção das séries temporais a serem utilizadas como entradas nos modelos de RNA.

O Capítulo 5 apresenta a análise dos resultados advindos do método aplicado a este trabalho, aplicando-se os modelos de RNA para classificação das séries temporais, bem como uma análise de suas potencialidades.

Finalmente, o Capítulo 6 resume as principais realizações deste trabalho, estabelece algumas diretrizes para ações futuras e apresenta também limitações encontradas ao longo do desenvolvimento deste trabalho.

2 Fundamentação Teórica

Ao longo deste Capítulo, serão apresentados alguns conceitos introdutórios referentes às tecnologias necessárias para a compreensão deste trabalho. Inicia-se apresentando os conceitos da tecnologia *blockchain* na Seção 2.1. Em especial, há um aprofundamento nos conceitos específicos da rede *Blokchain* Ethereum e seu processo de contratos inteligentes na Seção 2.2. Depois, na Seção 2.3, passa-se para uma descrição do aspecto econômico e mercadológico das criptomoedas e a explicação sobre os tipos de fraudes. Logo após, na Seção 2.4 são apresentados os principais conceitos sobre série temporais e modelos de previsibilidade nas mesmas. E, finalmente, na Seção 2.5, encontram-se abordados alguns conceitos pertinentes às RNA.

2.1 Tecnologias *Blockchain*

As tecnologias de *Blockchain* foram desenvolvidas com um núcleo comum, independente da criptomoeda envolvida (Bitcoin, Ethereum, IOTA, etc), de natureza criptográfica. Trata-se de uma confluência entre redes, sistemas distribuídos e criptografia para resolver o problema de intermediação de valor.

Entre os fundamentos dessa tecnologia, destaca-se, no seu núcleo, a função de *hash*, descrita na Figura 2.1. Esta função matemática, de maneira genérica, precisa apresentar três propriedades principais: a entrada precisa ser composta de uma sequência de caracteres de qualquer tamanho, a saída deve produzir um valor (*hash*) de tamanho fixo e deve também possuir eficiência computacional (O(n)).

Seu uso em criptomoedas também precisa definir outras propriedades igualmente importantes, a saber: resistência à colisão, que acontece quando dois valores de entrada distintos não produzem uma mesma saída, conforme a Figura 2.2; e ocultação, onde é impossível calcular o valor da entrada inicial a partir do valor da saída da função *hash*.

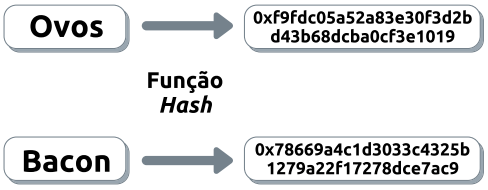


FIGURA 2.1 – Exemplo de função *hash*

FIGURA 2.2 – Exemplo de resistência à colisão

2.1.1 Arquitetura do Livro-Razão Público Distribuído

A tecnologia *Blockchain* se baseia no registro independente de blocos de transações, escritos via algoritmo distribuído, onde se situa o sistema, e compartilhados em uma rede par-par. Os blocos do livro-razão são organizados em cadeias de dados ou, como ilustrada na Figura 2.3, uma lista encadeada.

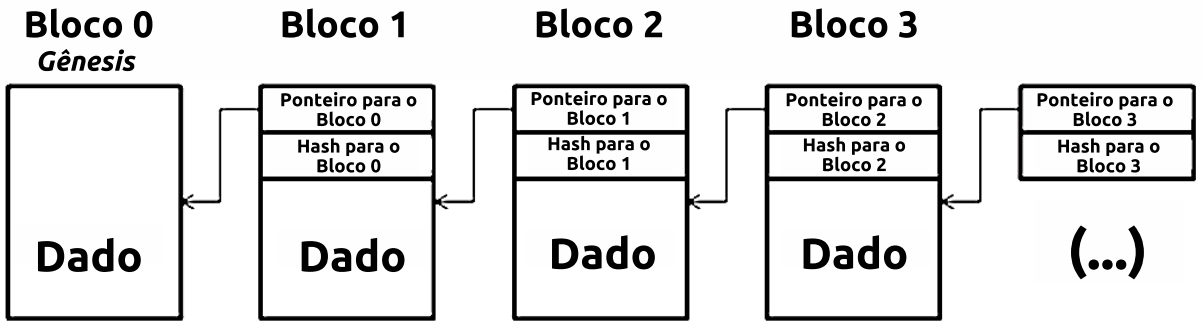


FIGURA 2.3 – Cadeia de blocos baseada em uma lista encadeada.

Lista encadeada é um tipo de estrutura de dados comum em computação, mas no caso da *blockchain*, os ponteiros foram substituídos por ponteiros *hash*. O ponteiro *hash* (NARAYANAN *et al.*, 2019) é um ponteiro que indica para onde determinado dado é armazenado junto com um *hash* criptográfico do próprio dado. Desta forma, não tem por finalidade somente recuperar informações, ao apontar para o bloco anterior, mas também é uma maneira de verificar se os dados não foram alterados, de acordo com a Figura 2.4.

O primeiro bloco da lista em *blockchain*, ou seja, o primeiro a ser criado, é chamado de Bloco Gênesis. (NARAYANAN *et al.*, 2019). Internamente no bloco, os dados são organizados como uma Árvore Merkle, que é uma estrutura de dados em árvore binária (DANNEN, 2019) composta de ponteiros *hash*, valores *hash* como nós e transações como folhas, conforme a Figura 2.5.

A árvore Merkle possui propriedades importantes como: Cada transação T(A) até T(H) é convertida em um valor *hash* H(A) até H(H). Desse modo os valores *hash* são pareados a fim de gerar um novo valor *hash*. Caso não seja uma árvore completa, o valor

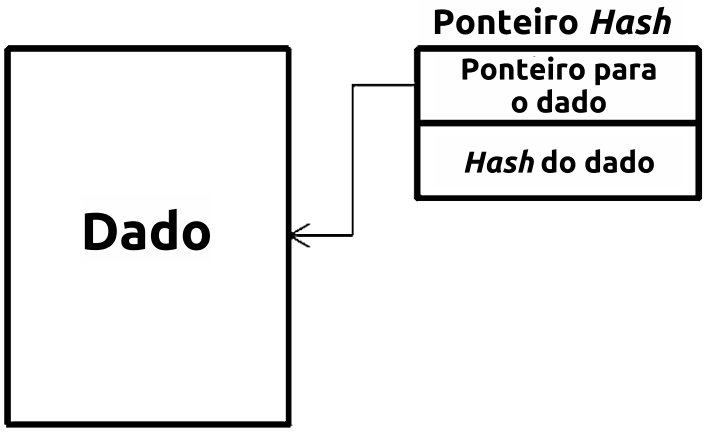


FIGURA 2.4 – Ponteiro *hash*.

hash é pareado consigo mesmo. Por exemplo, se o último elemento for H(G), então fica H(GG). E o processo então, continua até chegar à raiz da árvore. Este será o valor *hash* do bloco.

Tais propriedades criptográficas são embarcadas no programa que roda em cada nó da rede P2P, apresentada na Figura 2.6, formando a arquitetura de Livro-Razão Distribuído (NARAYANAN *et al.*, 2019). Entre suas propriedades, destaca-se a imutabilidade pois, para todo novo bloco, é preciso consistência e unicidade criptográfica com o bloco anterior.

2.1.2 Autenticidade das Transações

Seguindo as definições dessa tecnologia, no momento em que não há uma autoridade central reguladora, é necessário dispor de um meio que garanta a autenticidade ao longo

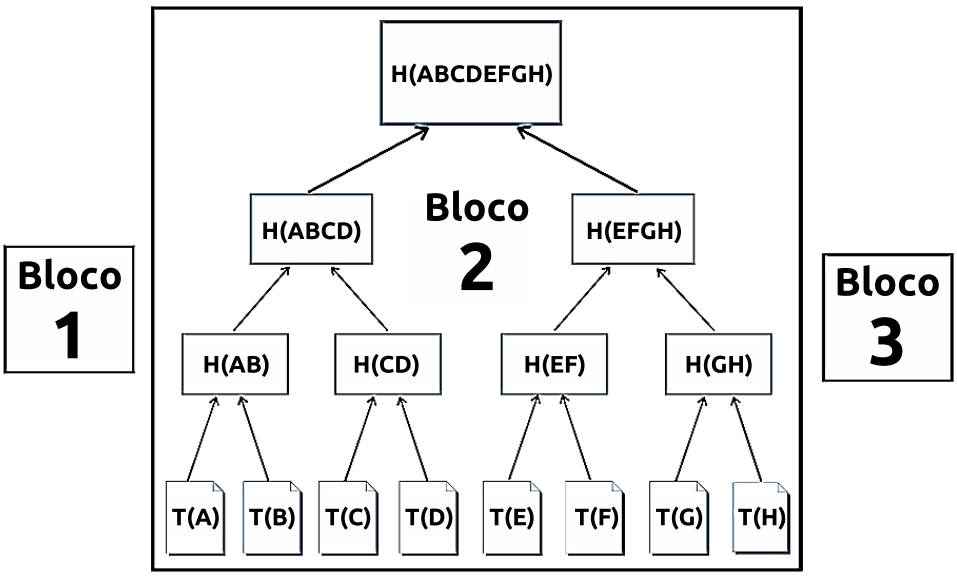


FIGURA 2.5 – Árvore Merkle.

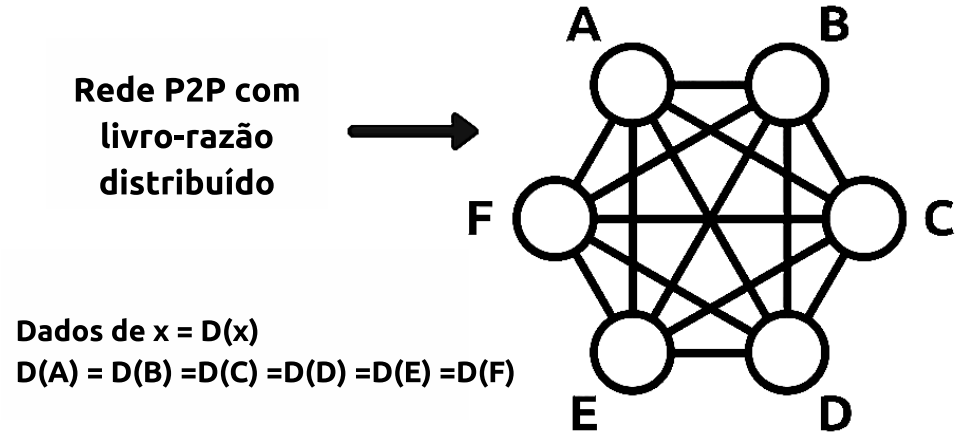


FIGURA 2.6 – Rede P2P com Livro Razão Distribuído.

das trocas de mensagens na rede. Portanto, a assinatura digital é um método de autenticação de informação digital, e é utilizada na cadeia de blocos com livro-razão distribuído, semelhante a uma assinatura. As assinaturas em *Blockchain* (NARAYANAN *et al.*, 2019) são feitas por meio de criptografia assimétrica com chaves públicas e privadas, criadas a partir de um número aleatório inicial e, posteriormente, com base nessas chaves, permitem assinar e verificar assinaturas, conforme Figura 2.7.

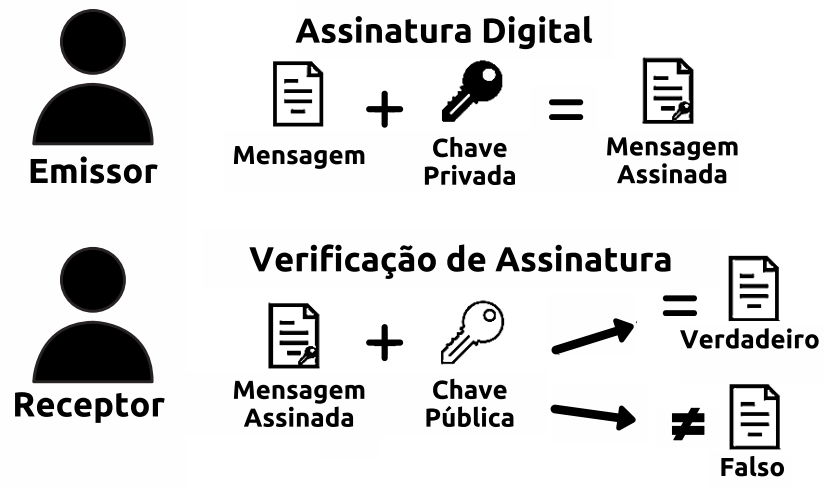


FIGURA 2.7 – Fluxo do Processo de envio e recebimento de mensagem com assinatura digital.

2.1.3 Algoritmo de Consenso

Caso se queira acrescentar algum bloco em uma estrutura de Livro-Razão Distribuído, é necessário que tal acréscimo seja replicado em todos os nós da rede, a fim de que todos os nós tenham o mesmo conjunto de dados. No entanto, neste tipo de estrutura, não há uma autoridade central para validar esta alteração.

Dessa forma, é fundamental a presença de um mecanismo de consenso, a fim de que todos os nós da rede cheguem a um acordo comum sobre o estado atual do Livro-Razão Distribuído. O consenso da rede é realizado por meio do algoritmo de consenso, o qual tem como propósito garantir a confiança entre pares desconhecidos.

Essencialmente, o algoritmo de consenso garante que cada novo bloco adicionado à cadeia de blocos seja a única versão verdadeira, que é acordada por todos os nós (LAURENCE, 2019). Prova de Trabalho (PoW) e Prova de Posse (PoS) são alguns exemplos de algoritmos de consenso e serão abordados mais adiante.

2.1.4 Aplicabilidade e Resumo das Principais Características

Como pode-se observar, *Blockchain* é resultado de uma combinação de arquitetura de rede P2P, criptografia (assinatura digital, funções hash, árvore Merkle, desafios criptográficos), banco de dados distribuído (livro-razão distribuído), algoritmos de consenso, dentre outros fatores (GREVE *et al.*, 2018).

A sua aplicabilidade traz muitas possiblidades, por exemplo, especificamente na área financeira, facilita as transferências de valor ponto a ponto de todos os tipos, de moeda digital a mercadorias físicas e títulos de propriedade, sem a necessidade de um intermediário, como bancos, contadores ou advogados. Portanto, está no centro de muitas perspectivas promissoras, destinadas a melhorar a eficiência, transparência e segurança em todos os tipos de negócios e transações sociais (FRIZZO-BARKER *et al.*, 2020).

Em resumo, pode-se destacar que as tecnologias *Blockchain* trazem novas propriedades interessantes para a construção de sistemas financeiros, como o financiamento coletivo. A descentralização é feita por meio do estabelecimento da confiança entre as partes, sem a necessidade de uma entidade central. A integridade é garantida porque os dados são replicados em todos os nós de maneira segura e confiável. Os blocos são rastreáveis e transparentes, pois são ordenados sequencialmente. Os usuários possuem uma garantia de privacidade, no sentido de que a gerência das chaves privadas é responsabilidade de cada um. As transações entre os usuários são realizadas de maneira descentralizada, sem a presença de uma autoridade central (XU *et al.*, 2016). E o consenso é realizado por meio de um algoritmo de consenso.

2.2 Ethereum

Em 2008, o Bitcoin foi criado e se tornou a primeira criptomoeda de grande adesão ao mercado (NAKAMOTO, 2019) e, como algoritmo de consenso, utiliza o PoW (*Proof of Work*). Nos anos seguintes, outras foram surgindo no mercado e, por sua vez, implemen-

tando novas tecnologias, como o Litecoin, em 2011, que modificou o protocolo do Bitcoin, de forma a aumentar a velocidade das transações. Outrossim, o Peercoin, em 2012, criou um sistema híbrido de mineração, baseado em PoW e PoS *Proof of Stake*.

A rede *Blockchain* Ethereum foi criada em 2014 por Vitalik Buterin (BUTERIN *et al.*, 2014). Comparativamente às outras, *blockchain* traz uma importante inovação que é suporte ao desenvolvimento rápido dos chamados *Smart Contracts*, que são contratos de execução automática (ABDELHAMID; HASSAN, 2019), armazenados na *blockchain* em forma de código binário, com o objetivo de permitir que partes trabalhem juntas, mesmo que não confiem umas nas outras. Desta forma, os *Smart Contracts* revolucionaram muitos setores que têm aplicações financeiras, tornando a tecnologia *Blockchain* cada vez mais relevante. (ZHAO *et al.*, 2016)

2.2.1 Aspectos Técnicos

Conforme seu sítio oficial[1], o Ethereum é pautado por princípios de simplicidade, universalidade, modularidade, agilidade e não à discriminação ou censura. A ideia, em resumo, é que o protocolo não introduza maior complexidade e se mantenha simples, que seja desenvolvido de maneira modular e que tenha um caminho de atualizações, fácil de ser implementado. Além disto, no quesito dos contratos inteligentes, que estes possam usar de uma linguagem de *script* interna de Máquina de Turing completa para qualquer aplicação. E, finalmente, que não sejam restringidos os usos da plataforma.

2.2.1.1 Contas Ethereum

Na rede Ethereum, os usuários interagem com a rede por meio das contas (*accounts*), que podem ser do tipo carteira (*wallet*) ou contratos (*contracts*) (SOLOLON, 2019). As carteiras são contas gerenciadas por humanos e não possuem códigos associadas a elas. Por meio delas, é possível realizar transações de uma carteira até outra conta, mediante o uso da chave privada do próprio usuário dono da carteira. Os dados mais importantes da carteira e de interesse no nosso estudo são:

- Endereço (*address*) — Cada conta possui apenas um único endereço, o qual é um *hash* hexadecimal de 42 caracteres, usado para identificar a conta;
- NONCE — um contador incremental único usado para garantir que cada transação possa ser realizada uma única vez;
- Saldo atual da conta (*balance*) — saldo em Ether, que é a criptomoeda nativa do Etherem e é utilizada para pagar qualquer taxa de transação; e

[1] https://ethereum.org/en/whitepaper/

- Armazenamento da conta — vazio por padrão.

Por outro lado, os contratos inteligentes são contas não gerenciadas por humanos. Todas as vezes que elas recebem uma transação com alguma mensagem, seus códigos internos são ativados, permitindo que elas operem de forma autônoma, conforme a lógica dos seus próprios códigos. Uma vez executado, o registro das ações é imutável, pois ele se encontra registrado dentro da rede Ethereum. No *smart contract* existem, além das quatro informações anteriores (endereço, NONCE, saldo atual da conta e armazenamento da conta), o acréscimo de mais uma:

- Código do contrato — contém a lógica a ser executada, quando um determinado evento acontecer (como alguém enviando mensagens para o contrato).

2.2.2 Transações e Mensagens no Ethereum

Na rede Ethereum, as contas se comunicam, mudando o seu estado, por meio de mensagens e transações. Quando a comunicação é vinda de uma carteira, é chamada de transação. Por outro lado, quando é vinda de um contrato, é chamada de mensagem.

Todas as transações e mensagens contém uma ou mais instruções. Cada instrução a ser executada possui um custo medido pela unidade valorada chamada GAS. A quantidade de GAS para execução da instrução é estabelecida pelo Ethereum *yellowpaper*[2], um documento que contém todas as especificações técnicas da rede Ethereum, de acordo com o tamanho e complexidade do código da instrução.

Conforme o Ethereum *yellowpaper*, o custeamento das transações obedece à seguinte lógica:

1. O usuário envia à rede a transação e uma quantidade de dinheiro chamada GASLIMIT (GAS Limite), que é a quantidade de GAS máxima que o usuário aposta ser necessária para a execução da sua transação;

2. A rede, por sua vez, executa a operação, reportando o valor nominal, cujo nome é GASUSED, (GAS usado) para a execução da transação; e

3. Caso o GASUSED seja menor do que o GASLIMIT, como por exemplo o caso 1 da Figura 2.8, a rede devolve o troco em GAS que o usuário enviou, mas não foi utilizado. No entanto, se o GASUSED for igual ou superior ao GASLIMIT, a operação é revertida para a situação inicial e, consequentemente, a rede não devolve o GAS ao usuário, entrando no caso 2 da Figura 2.8.

[2]https://ethereum.github.io/yellowpaper/paper.pdf

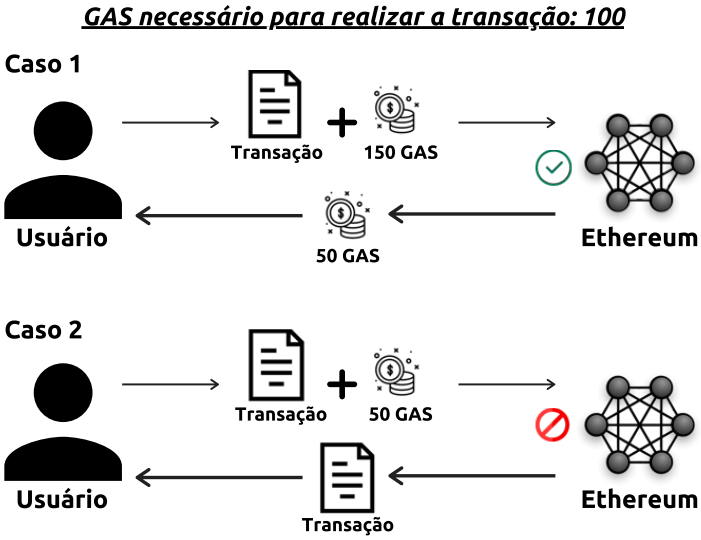


FIGURA 2.8 – Fluxo do Processo de transação e GAS usado.

As transações e mensagens contém os seguintes campos:

- *Hash* da transação, número hexadecimal composto de 66 caracteres,
- Origem da mensagem,
- Destinatário da mensagem,
- *NONCE* da mensagem do remetente,
- Assinatura identificando o remetente (somente as transações possuem),
- Número do bloco que a transação pertence,
- Quantidade de Ether a ser transferida do remetente para o destinatário,
- Campo de dados opcional,
- GASLIMIT, representando a quantidade máxima de GAS que o usuário está disposto a gastar em uma determinada transação,
- GASUSED, indicando o total de GAS utilizado para a realização da transação e
- Preço do GAS (GASPRICE), representando a taxa que o usuário remetente está disposto a pagar por cada etapa computacional.

Outro aspecto importante é a *Ethereum Virtual Machine*, que é uma máquina de Turing quase completa[3] onde são executadas todas as transações. O termo "quase" é colocado devido ao fato de que o cálculo é intrinsecamente limitado por um parâmetro, o

[3]https://ethereum.github.io/yellowpaper/paper.pdf

GAS, que limita a quantidade total de computação feita. A linguagem de programação em que é feita a interação com a EVM, ou seja, que são desenvolvidas as aplicações em Ethereum se chama Solidity, cuja sintaxe é semelhante ao do javascript.[4]

2.2.3 Contratos Inteligentes

No contexto desta pesquisa, é preciso deixar claro que os processos das atividades de ICO são programados dentro de um contrato inteligente. Portanto, outros detalhes de contrato inteligente são pertinentes. O contrato inteligente é um programa executável, imutável, escrito em uma conta contrato, na *blockchain* Ethereum (FARELL, 2015), que executa automaticamente uma transação, quando certas condições predeterminadas são atendidas ou um evento específico identificado é acionado. Seu código fonte é aberto (basicamente um bytecode irreversível na EVM) e, desta forma, torna-se transparente para todos, possibilitando que seja criado um contrato de confiança entre diversas partes, sem a necessidade de uma organização central (ANTONOPOULOS; WOOD, 2018).

Adicionalmente, dentro do processo de atividade de ICO, o contrato inteligente é utilizado para cunhar o que são conhecidos como *tokens*, que são novas criptomoedas, representando ativos digitais, os quais podem ter o valor de parte de uma empresa (como se fosse uma ação) ou até mesmo possibilitar o acesso a um determinado serviço.

O evento em que um contrato inteligente se comunica com outro é chamado de transação interna, o qual não é registrado na rede Ethereum. Caso se queira visualizá-la, é preciso executar a transação e rastrear as chamadas que ela fez.

Além de atividades de ICO, o advento dos contratos inteligentes tornou possível desenvolver quaisquer aplicativos rodando dentro da *blockchain*, conhecidos como Aplicativos Descentralizados (DApps). A Figura 2.9 mostra como é feita a comunicação entre pessoas que estejam utilizando um DApp. A lógica da aplicação não está em sua camada externa, mas sim, armazenada na *blockchain*. Desta forma, há o princípio de que todo o processo seja exposto ao usuário de uma forma transparente.

Também existem outros princípios (MUKHOPADHYAY, 2018) como: o aplicativo não deve ter um único ponto de falha, os registros de operação são armazenados criptograficamente em um local público *blockchain*, bem como o aplicativo deve usar algum tipo de *token* nativo para vários processos e acesso, e qualquer contribuição de valor deve ser recompensada em termos de tais *tokens*. Finalmente, o aplicativo deve usar algum padrão criptográfico aceito como prova do valor, contribuído por um nó, para gerar *tokens*.

[4]https://solidity.readthedocs.io/en/v0.7.1/

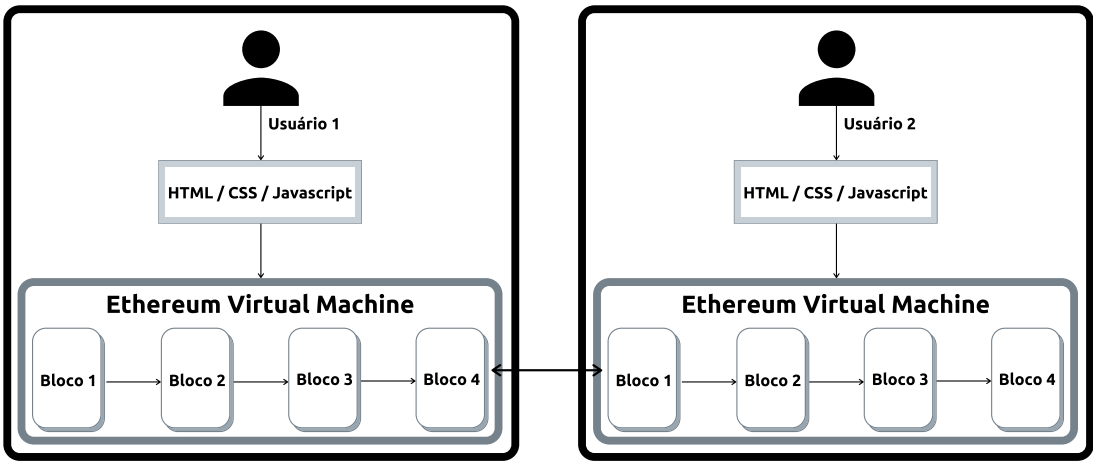


FIGURA 2.9 – Funcionamento de um DApp.

2.3 Aspectos Econômicos de Criptomoedas

Esta Seção apresenta uma introdução do uso de plataformas com tecnologia *Blockchain* como ativos financeiros digitais, chamadas de criptomoedas. Estes ativos permitem a posse e transferência de "reservas de valor" entre partes, garantidas por meio de uma tecnologia criptográfica descentralizada, presente na *blockchain*, ao invés de um banco ou qualquer presença de terceiros (GIUDICI *et al.*, 2020).

Além disso, criptomoedas podem ser definidas como um sistema que atenda algumas condições para se tornarem viáveis como moedas digitais, conforme (LANSKY, 2019). Por exemplo, elas não precisam de autoridade central, o sistema deve preservar as unidades de criptomoeda atreladas aos seus respectivos proprietários, bem como o sistema deve definir se novas unidades de criptomoeda podem ou não ser criadas. Há outros aspectos de validação, como a posse de cada unidade de criptomoeda poder ser provada, exclusivamente, por meio da criptografia, e transações poderem somente ser realizadas mediante a autorização do seu respectivo proprietário.

Diante desse cenário, o crescimento das popularidade das criptomoedas como ativos econômicos se deu por conta dos seguintes aspectos: o avanço da eficiência dos algoritmos de criptografia e a vontade de se criar uma moeda livre e autônoma, que não tivesse influência ou controle de governos, e também, que a reserva de valor fosse imune à inflação (ULRICH, 2017).

2.3.1 Oferta Inicial de Moedas (*Initial Coin Offering* - ICO)

Dentre os aspectos econômicos, a possibilidade de financiamento coletivo atraiu muitos usuários para essas plataforma. Em especial, conforme já foi explicado, atividade de ICO é um método de financiamento coletivo utilizado por diversas *Startups*, a fim de levantar

capital para a execução dos seus projetos de negócio (CONLEY, 2017). É composto por três fases (HAHN; WONS, 2018):

1. Planejamento (Pré - ICO) — Nesta fase, a empresa coloca na *blockchain* pública os seus *tokens*, que representam ativos do próprio empreendimento (semelhante às ações em um IPO), e o *Smart Contract*, que contem todas as regras de compra e venda destes ativos. Os *tokens* podem apresentar diversas funções como: criptomoeda, acesso a algum serviço, direitos de governança, direitos de lucros ou até mesmo, direito de contribuição de código (ADHAMI *et al.*, 2018). Adicionalmente, a empresa elabora o *White Paper* (QUEST, 2018) do seu *token*, um documento disponibilizado ao público, por meio do sítio do próprio projeto, com o objetivo de convencer as pessoas a investirem no seu empreendimento. Contém termos e condições para se fazer parte do projeto, valor necessário em dinheiro para se iniciar o projeto, descrição da equipe envolvida no projeto, detalhes técnicos do projeto e demais informações relevantes, como funcionalidade dos *tokens*. Contudo, tal documento não possui qualquer validade jurídica, ou seja, a confiança mútua entre investidor e empreendedor é a base do financiamento do negócio;

2. Evento de Lançamento — Nesta fase, conforme a Figura 2.10, os *tokens* são colocados à venda. As pessoas interessadas começam a comprá-los, mediante suas próprias análises de viabilidade do projeto. Os interessados fornecem uma quantia de criptomoedas (já consolidada no mercado, por exemplo, usando Ethereum ou Bitcoin) e, em troca, os compradores recebem o valor equivalente em *tokens*. As criptomoedas que os interessados deram vão para as empresas, a fim de que elas possam trocá-las por moeda corrente (como o Dólar, por exemplo) e, desta forma, possuir capital para executar seus projetos; e

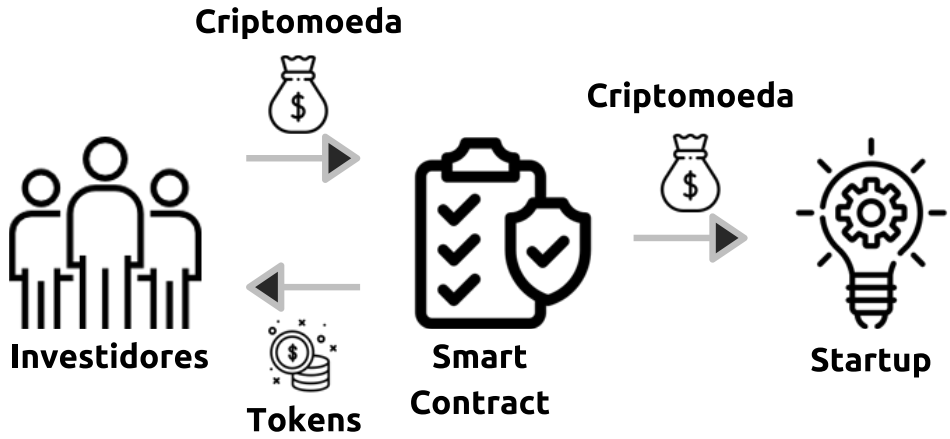


FIGURA 2.10 – Fluxo de aquisição de tokens em um ICO.

3. Pós - ICO — Por último, nesta fase, caso a *startup* consiga recursos necessários para começar seu empreendimento, ela dá início ao seu projeto. Futuramente, se

a empresa se valorizar no mercado, os seus *tokens*, por sua vez, também irão se valorizar, até chegar ao ponto em que casas de câmbio (*exchanges*) os aceitem como uma criptomoeda. Especuladores investem na compra de *tokens* de empresas que ainda não lançaram seus projetos, com a esperança de que ela se valorize no mercado.

Sendo assim, o advento das atividades de ICO contribuiu para o aumento do número de criptomoedas no mercado (KHER *et al.*, 2020). Esta popularização aconteceu no mundo todo. Em especial, países com menos regulações sobre ICOs são mais atrativos. Estados Unidos, Rússia, Reino Unido, Cingapura e Suíça lideram o ranking de países com mais ICOs, conforme a Figura 2.11.

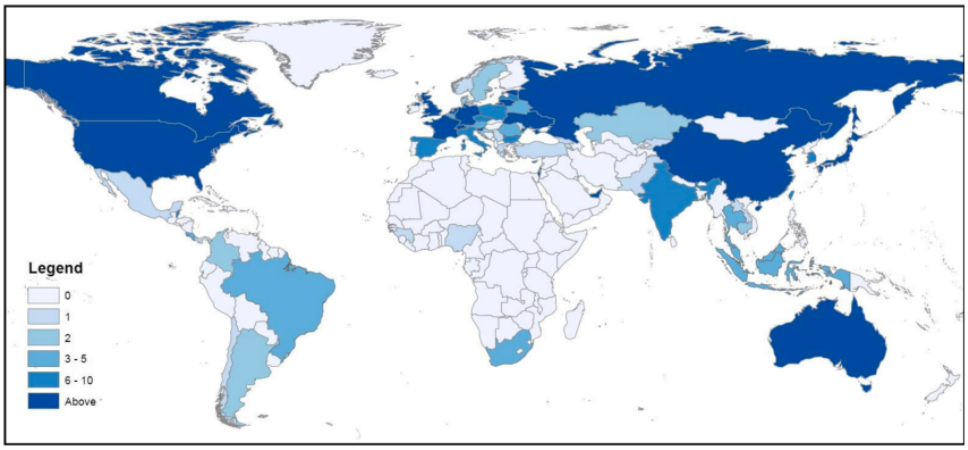


FIGURA 2.11 – Mapa mundi com o número de ICO por países.

2.3.2 Outros Conceitos do Mercado de Criptomoedas

Depois do processo de ICO concluído, os *tokens* se tornam novas criptomoedas, e sua valorização acontece com o tempo. O valor da nova moeda vai gerando maior valor em função da sua procura.

2.3.2.1 Capitalização de Mercado

Primeiramente, é preciso entender dois conceitos em mercado monetários: *Circulating Supply* e *Total Supply*. O primeiro se refere ao total de moedas circulando no mercado e pode crescer ao longo do tempo, na medida em que novas moedas mineradas entrarem no mercado (GIUDICI *et al.*, 2020). Por outro lado, o segundo se refere ao total de moedas que poderá existir, incluindo aquelas que ainda não foram mineradas. Multiplicando-se o *Circulating Supply* pelo preço atual é possível obter o valor de Capitalização de Mercado (*Market Cap*) da criptomoeda. Normalmente, no mercado de criptomoedas, os sítios que

provêm informações de criptomoedas utilizam este dado como um meio de ranqueamento, como o coinmarketcap[5] e o coingecko[6].

2.3.2.2 *Exchanges*, Carteiras e Titulares

Com a capitalização expressiva do mercado de criptomoedas, o fenômeno contribuiu para o surgimento das *exchanges*, equivalentes às casas de câmbio de criptomoedas (LAURENCE, 2019). Desta forma, qualquer proprietário de criptomoedas, incluindo *tokens*, pode abrir contas e negociá-los por outras criptomoedas ou por moedas reguladas pelo governo (chamadas de FIAT - em latim "let it be done").

Os proprietários de algum ativo de uma criptomoeda são chamados de titulares (*holders*). Os proprietários podem ser carteiras ou contratos inteligentes. Com relação às carteiras, elas podem ser classificadas em: carteira quente (*hot wallet*) ou carteira fria (*cold wallet*) (LANSKY, 2019). A primeira se refere a uma carteira constantemente conectada à internet, sob posse de uma *exchange*, enquanto a segunda, não fica conectada integralmente à internet, tendo como posse o próprio usuário.

No caso das *exchanges*, há dois tipos básicos: as centralizadas e as descentralizadas (DEX). As centralizadas são empresas que possuem grande repositório de valor na *blockchain* para negociar diretamente entre cada usuário as trocas em sua própria plataforma. As *exchanges* descentralizadas permitem a troca direta entre proprietários. Frequentemente, DEXs são projetadas de forma totalmente integradas à estrutura da tecnologia *Blockchain*. Isto aumenta a segurança, mas as torna dependentes da velocidade da inserção do bloco à rede. Sua velocidade lenta e os altos custos de transação que vêm com a integração total também as tornam pouco atraentes para os comerciantes.

2.3.3 Fraudes

Desde que o Bitcoin inaugurou o mercado de criptomoedas, a partir de 2009, a livre troca descentralizada de recursos passou a ser um atrativo para o engajamento nessas plataformas. No entanto, os fraudadores evoluíram com o tempo, encontrando novas maneiras de enganar as pessoas. Sobre as fraudes neste mercado, existem quatro tipos, segundo (BAUM, 2018):

- Esquema Ponzi é uma operação de investimento fraudulenta em que o falso empreendedor gera retornos para investidores mais antigos, por meio da receita paga por novos investidores, ao invés de atividades comerciais legítimas ou lucros de negociação financeira (CHEN *et al.*, 2018);

[5]https://coinmarketcap.com/
[6]https://www.coingecko.com

- Esquema de Fuga é uma prática fraudulenta de promotores de criptomoedas que desaparecem com o dinheiro dos investidores durante ou após um ICO. Às vezes, o empreendimento pode ter começado como uma empresa legítima. Todavia, devido a fatores econômicos adversos, mau planejamento de negócios ou uma combinação de ambos, os responsáveis pelo empreendimento podem desaparecer, numa tentativa de escapar das possíveis consequências de uma falência;

- *Pump and Dumps* é um tipo de fraude que consiste em tentar aumentar o preço de uma criptomoeda, com base em notícias enganosas. Desta forma, o empreendedor malicioso compra uma quantidade considerável de criptomoedas a um preço normal para, após a liberação das *fake news*, vender por um preço maior (XU; LIVSHITS, 2019); e

- Há o furto de criptomoedas, que pode ser executado por meio de falsas carteiras (*wallets*, *emails*, sítios ou aplicativos para celular. Além disto, pode haver o furto de dados, atividade conhecida como "*Phishing*".

2.4 Séries Temporais

Neste trabalho, são abordados modelos de classificação utilizando RNA, a partir de dados de séries temporais. Portanto, uma revisão sobre o assunto é adequada para o seu entendimento.

2.4.1 Definições de Séries Temporais

A Série Temporal é o conjunto de observações, representadas por valores numéricos, igualmente espaçados ao longo do tempo (MILLS, 2019). O conceito é usado para analisar o comportamento de uma determinada variável ao longo do tempo. O eixo Y representa os valores da variável e o eixo X, o tempo igualmente espaçado. São amplamente utilizadas em diversas áreas, como: índice da bolsa de valores, registro meteorológico periódico de uma cidade, taxa de natalidade ao longo do tempo; dentre outras aplicações.

As séries temporais apresentam os seguintes componentes (PAL; PRAKASH, 2017) básicos, conforme a Figura 2.12[7]: Tendência (padrão de crescimento), Sazonalidade (repetição periódica do padrão), Ciclo (sazonalidade mais espaçada) e Ruído (flutuação irregular).

[7]http://www.abepro.org.br/biblioteca/enegep2010_tn_stp_115_755_16795.pdf

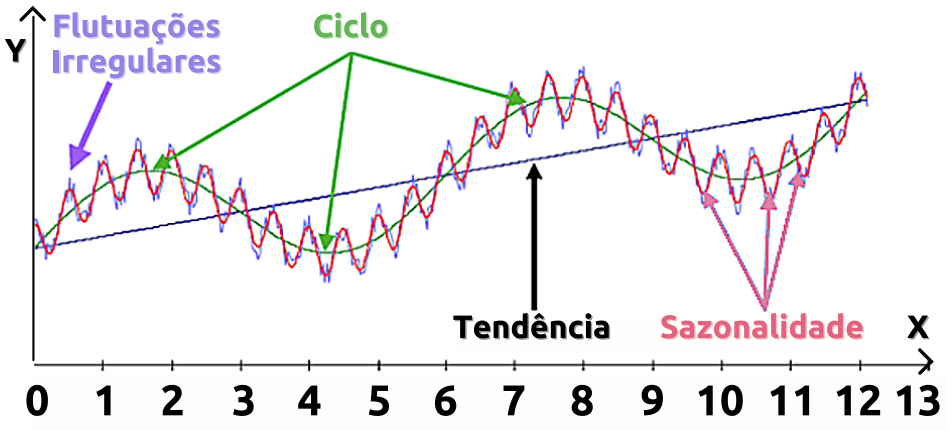


FIGURA 2.12 – Componentes de uma série temporal.

2.4.2 Autocorrelação

Dentre os parâmetros importantes de séries temporais, a função de Autocorrelação (Equação 2.1) é a dependência que o valor da variável possui em relação as suas observações passadas (BOX *et al.*, 2011). Dadas as medições, Y_1, Y_2, ..., Y_n, no tempo X_1, X_2, ..., X_n, a função de autocorrelação de atraso k é definida como:

$$r_k = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{2.1}$$

Embora a variável de tempo, X, não seja usada na fórmula, a suposição é que as observações sejam igualmente espaçadas. A saída da função é o coeficiente de correlação. No entanto, em vez de correlação entre duas variáveis diferentes, a correlação é entre dois valores da mesma variável nas vezes X_i e X_i + k. A autocorrelação tem como utilidade identificar se a série possui tendência, períodos (sazonalidade ou ciclos) ou se não possui nenhuma das duas anteriores (conhecido como Ruído Branco (*white noise*).

Os valores de saída (eixo y) estão dentro do intervalo de -1 e 1, sendo que, quanto mais próximo de zero, menos autocorrelação haverá. Portanto, uma independência estatística entre amostras. Adicionalmente, se o valor for maior que |0.2|, indica uma autocorrelação significativa. Os valores de entrada (eixo x) variam de acordo com o número de observações passadas, em relação ao tempo de medição atual do valor da variável, conforme a Figura 2.13, onde a linha vermelha representa o valor de |0.2|.

2.4.3 Normalização dos dados

Outro aspecto que tem como o objetivo realizar a comparação efetiva entre duas séries, evitando que se tenha um viés para as variáveis de maior grandeza, é a normalização de

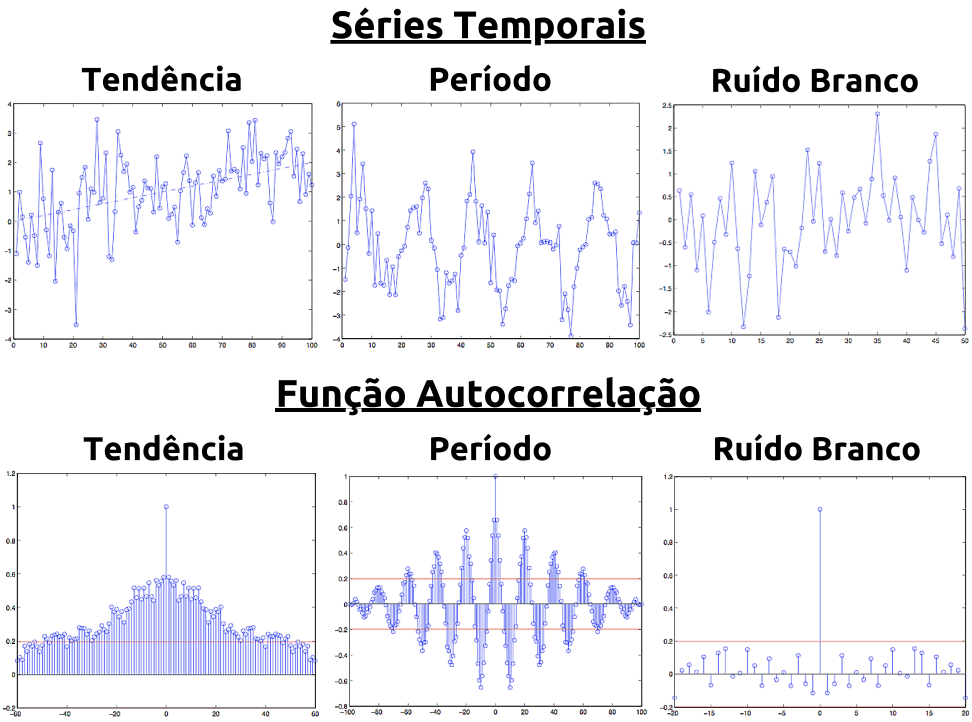


FIGURA 2.13 – Aplicação da função Autocorrelação em três séries temporais diferentes (HUANG *et al.*, 2020).

dados. Trata-se de processo de transformar os valores das variáveis em um intervalo escalável entre 0 e 1 (MILLS, 2019). O tipo de Normalização mais utilizada é a **min-max** (Equação 2.2). O cálculo é feito, a partir dos valores de Y_{min} e Y_{max}, conforme a Equação 2.2. Entretanto, não é tão eficiente quando a série possui valores *outliers*, com valores extremos de máximo e mínimo, onde uma normalização gera valores tendendo a zero.

$$Y_{norm} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \tag{2.2}$$

2.4.4 Segmentação

O último item de revisão em séries temporais é a segmentação. Trata-se do processo que envolve a divisão da série temporal em faixas de tempo, chamadas segmentos, apropriadas para a análise do problema a ser estudado (BANOS *et al.*, 2014). Conforme a Figura 2.14, a primeira técnica de segmentação é a simples, que consiste em fatiar segmentos específicos da série temporal para análise. Se por um lado ela é simples em adotar, por outro, dependendo do tamanho da amostra, pode haver perdas de informações relevantes para o problema em questão.

A segunda técnica é a de **janela deslizante** (KEOGH *et al.*, 2004), amplamente adotada em Reconhecimento de Atividade Humana por meio do uso de acelerômetros, os quais detectam atividades estáticas (sentar ou deitar) e dinâmicas (correr ou andar). A série

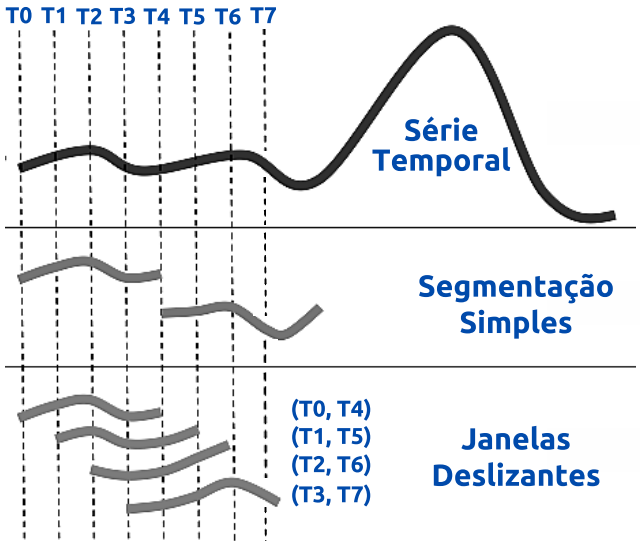


FIGURA 2.14 – Segmentação Simples e Janelas Deslizantes

temporal é separada em segmentos de tamanhos iguais (tamanho da janela), cujo o início e o fim recebem o incremento de $t + 1$. Esta técnica de segmentação é a base de como transformar o conjunto de dados de uma série em um problema de aprendizado supervisionado (BROWNLEE, 2018). Se o tamanho da janela for pequeno, permite a detecção de atividades mais rapidamente. Contudo, se for grande, possibilita o reconhecimento de atividades mais complexas, embora seja necessário um tempo maior para processar o segmento.

2.5 Redes Neurais Artificiais

A última parte relevante de conhecimentos fundamentais é a revisão de modelos de RNA aplicados neste trabalho para a detecção de fraude em criptomoedas advindas de atividades de ICO.

Neste trabalho, foram utilizados modelos baseados em RNA, em detrimento de outros modelos mais clássicos como ARIMA, devido ao fato de realizar a classificação e não puramente previsão de dados. Além disto, as características complexas presentes nas transações na rede *Blockchain* Ethereum não são apropriadas para modelar rajadas de transações e forte variação de dados, usando modelos mais simples. Portanto, os modelos baseados em RNA, que possuem maior poder de detecção de padrões, são mais apropriadas neste contexto. Modelos de redes neurais são mais genéricos que os tradicionais modelos ARIMA, que embarcam somente aspectos de sazonalidade, tendência e variância.

2.5.1 Definição de Redes Neurais Artificiais

Em termos de conceituação, as RNA tem a sua origem inspirada nos estudos sobre a estrutura das redes neurais biológicas, em particular o cérebro humano, de forma que seja possível um sistema computacional resolver problemas que envolvam aprendizado e generalização de informações (ZURADA, 1992). Elas possuem diversas aplicações, como: reconhecimento de padrões (voz, texto, imagem, dentre outros), tanto previsão quanto classificação de séries temporais, suporte à tomada de decisão, otimização da qualidade de um determinado produto, desenvolvimento da robótica e etc. Neste trabalho, utiliza-se 3 tipos de arquiteturas que compõem os modelos de classificação: *Multilayer Perceptron* (MLP), Redes Neurais Convolucionais (CNN) e *Long - Short Term Memory* (LSTM).

2.5.2 Estrutura do Neurônio Artificial

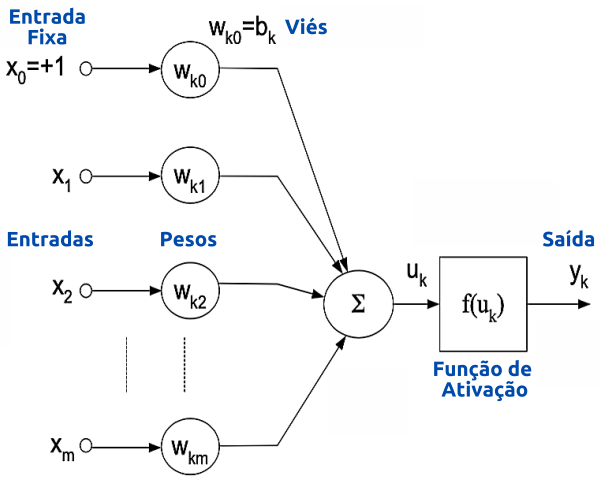


FIGURA 2.15 – Esquema de funcionamento de um neurônio artificial.

Uma RNA é composta por uma série de neurônios, que são, por sua vez, a unidade elementar de processamento de informações. Cada neurônio possui a seguinte configuração, descrita na Figura 2.15. Cada neurônio pode ter uma ou mais Entradas, que são os valores a serem processados. Além disso, cada valor de entrada é multiplicado por um Peso relativo à entrada. Os valores resultantes de Viés (*bias*) constante entram como parcela da soma da Junção Somadora. E o somatório de todos os valores de entrada, multiplicados pelos seus respectivos pesos, mais o viés, são enviados à Função de Ativação. Essa tem como objetivo normalizar a saída para o intervalo [0,1] ou, alternativamente, para [-1,1], a fim de restringir a amplitude de saída do neurônio. Além disso, introduz a não-linearidade ao modelo. Finalmente, a Saída é o valor resultante da Função de Ativação. Dessa forma,

a saída obedece a seguinte Equação 2.3:

$$y_k = f(u_k) = f\left(\sum_{j=1}^{m} w_{kj}.x_j + b_k\right) \tag{2.3}$$

Ou, considerando o viés como entrada de valor x_0=1 e peso $w_{k0} = b_k$:

$$y_k = f(u_k) = f\left(\sum_{j=0}^{m} w_{kj}.x_j\right) \tag{2.4}$$

2.5.3 Funções de Ativação

No trabalho, conforme a modelagem, foram experimentados vários tipos de função de ativação (HAYKIN *et al.*, 2009), por exemplo, a função Sigmóide, que é muito usada para problemas que envolvam classificação e a Unidade Linear Retificada (ReLU), que é amplamente utilizada para resolução de vários problemas atualmente. Maiores detalhes destas funções são apresentadas nas Figuras 2.16 e 2.17, bem como na Figura 2.18 (BROWNLEE, 2018).

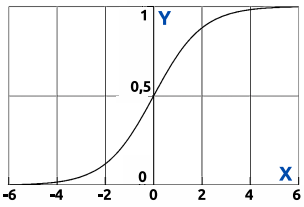


FIGURA 2.16 – Gráfico da função Sigmóide.

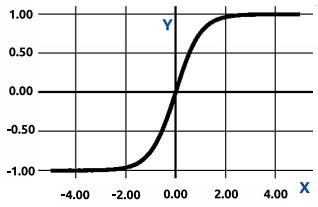


FIGURA 2.17 – Gráfico da função TANH.

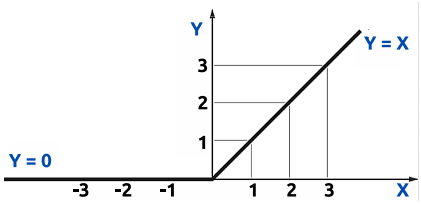


FIGURA 2.18 – Gráfico da função ReLU.

2.5.4 Arquiteturas

As RNAs podem ser analisadas sob uma ótica de um grafo orientado, no qual cada neurônio é equivalente a um vértice, e a direção do grafo corresponde ao caminho que as informações percorrem ao longo da rede. O conjunto de neurônios que possuem o mesmo grau é chamado de camada, que podem ser Camada de Entrada, Camada de Saída e Camada Escondida (MEHROTRA *et al.*, 1997). A densidade das camadas escondidas é que dão o nome de Redes Neurais Profundas.

Por outro lado, a orientação do grafo define as arquiteturas em: Redes Alimentadas Adiante (*Feed-forward*) e Redes Recorrentes. As Redes Alimentadas Adiante (*Feed-forward*) são redes onde não há ciclos. Podem ser de uma camada ou de várias camadas (redes profundas ou *deep learning*), conforme Figura 2.19. Enquanto as Redes Recorren-

tes são onde há ciclo, no qual a saída de um neurônio é aplicada como entrada no mesmo neurônio (Figura 2.20). Tal processo é chamado de realimentação e contribuem para o aumento da capacidade de aprendizagem da RNA, uma vez que as informações persistem ao longo do tempo. São as arquiteturas que mais se aproximam das redes neurais biológicas.

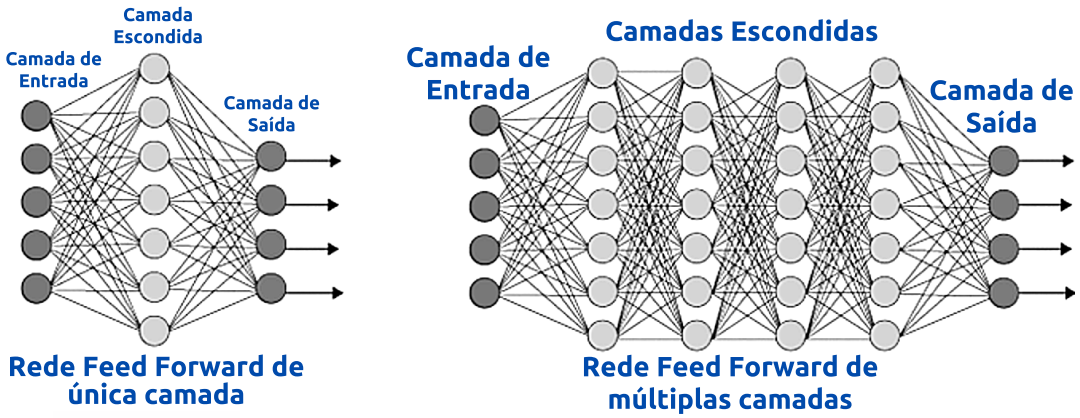


FIGURA 2.19 – Esquema de funcionamento de uma Rede Neural *feed-forward*.

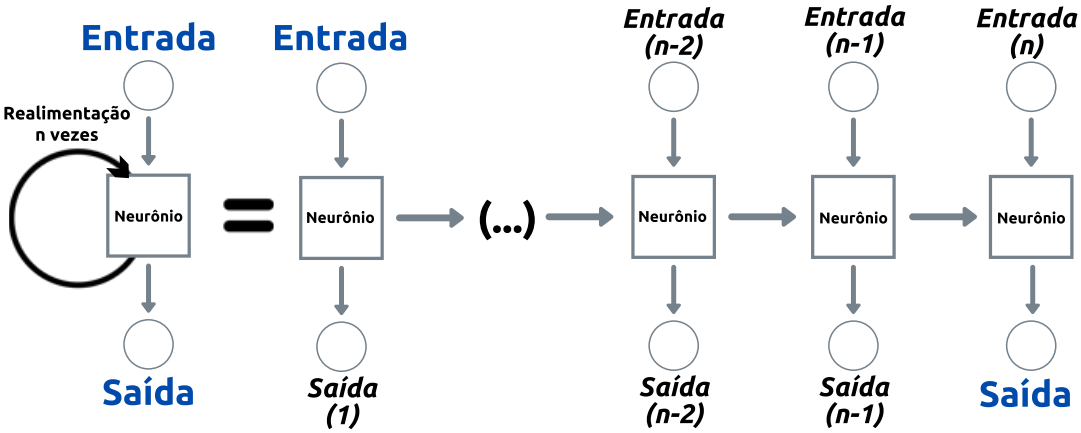


FIGURA 2.20 – Esquema de funcionamento de uma Rede Neural Recorrente.

2.5.5 Fase de Treinamento

O primeiro passo para que uma RNA possa resolver um determinado problema é efetuar o seu treinamento, um período no qual a RNA aprende informações relevantes de padrões de informações, de modo que ela atinja uma solução generalizada para uma classe de problemas. Exemplificando, supondo que o problema seja reconhecer um gato, a partir de uma imagem apresentada. O treinamento é o período em que ela aprende a reconhecer quais são os padrões da imagem de um gato.

Com relação ao **aprendizado** da RNA, assim como a classificação de imagens, também as séries temporais podem ser ensinadas para os modelos de RNA. Este aprendizado pode ser classificado em dois tipos (GOODFELLOW *et al.*, 2016):

- **Supervisionado** — A RNA dispõe de um conjunto de entradas e saídas esperadas, que serão a base do aprendizado. Dessa forma, é realizado um processo iterativo de ajustes de peso dos neurônios, a fim de que o valor de saída atual da RNA seja o mais próximo possível do valor de saída esperado, a partir de uma mesma entrada. O aprendizado supervisionado será o escopo deste trabalho; e
- **Não supervisionado** — Não necessita de um conjunto de entradas e saídas esperadas. A RNA, por si mesma, já realiza os ajustes dos pesos dos neurônios. O treinamento não supervisionado é normalmente usado para problemas de estimação, como clusterização, filtragem e distribuição estatística.

2.5.6 Algoritmo de Aprendizado

O algoritmo de aprendizado é um conjunto de regras bem definidas para a solução de um problema de aprendizado. Há vários tipos de algoritmos específicos para determinados modelos de redes neurais e eles, por sua vez, diferem entre si, principalmente pelo modo como os pesos são modificados. O ***Backpropagation*** (MEHROTRA *et al.*, 1997) será o algoritmo do escopo deste estudo. A escolha se deu por ser o algoritmo supervisionado mais utilizado. Ele consiste em duas fases:

- ***Forward*** — As informações pela rede, no sentido da camada de entrada até a de saída. Os pesos de cada neurônio são inicializados com valores aleatórios. Quando os valores chegam na camada de saída, eles são comparados com os valores esperados e, por sua vez, é calculado o erro, conforme a Equação 2.5.

$$E = \frac{1}{2}\sum_{i=1}^{p}(o_i - t_i)^2 \tag{2.5}$$

  Onde "p" é a quantidade de padrões, "o" é a saída atual e "t" é a saída desejada; e

- ***Backward*** — Após os erros serem calculados, os pesos dos neurônios da rede são atualizados no sentido camada de saída até a de entrada, conforme a Equação 2.6. Uma Taxa de Aprendizagem é adicionada, que indica a rapidez com que o vetor de pesos será atualizado. Caso seja muito pequena, o processo de treinamento se torna lento. Por outro lado, se for muito elevada, não será eficiente, uma vez que a variação dos valores de atualização dos pesos será alta.

$$\Delta w_i = -\gamma\frac{\partial E}{\partial w_i} \tag{2.6}$$

  Onde "w_i" é o peso do neurônio com índice "i", γ é a Taxa de Aprendizagem, que é multiplicada pela derivada do gradiente descendente (E).

Após os pesos serem modificados, o algoritmo passa para a fase *forward* novamente. E assim continua, iterativamente, com a finalidade de que o gradiente descendente seja igual a 0. Além disso, há ainda dois conceitos importantes sobre o treinamento de uma RNA (BRAGA, 2000):

- **Época** — Cada etapa de treinamento da rede é chamada de Época. Em geral, o treinamento acaba quando há o número de épocas é executado ou quando há uma convergência para o resultado esperado; e
- **Tamanho do *batch*** — É o número de amostras que serão treinadas em cada época. Se a rede treinar todas as amostras a cada época, pode ser que a eficiência não seja tão boa. Todavia, se for escolhido um número de amostras muito pequeno, pode haver a possibilidade de o número de amostras não refletir o padrão de todos que entrarão na rede, diminuindo a sua eficácia.

2.5.7 Fase de Teste e Métricas de Desempenho

É o período destinado a verificar se a RNA aprendeu a como solucionar o problema proposto. Basicamente, é verificar a eficiência do treinamento. Um aspecto importante é que não se pode utilizar quaisquer dados de teste utilizados previamente na fase de treinamento, a fim de que seja simulada uma situação o mais próximo da realidade possível. Portanto, os dados de aprendizado são separados em dados de treinamento e dados de teste.

Com relação aos erros, existem várias métricas para analisá-lo, como o erro médio absoluto (MAE) e o erro médio quadrático (MSE). Se a taxa de erros acima do esperado, é preciso verificar se houve algum problema durante o treinamento. Dois possíveis problemas são *underfitting* e *overfitting* (BRAGA, 2000).

Se o número de épocas for muito pequeno, ocasionará o *underfitting*, que é uma situação em que a rede aprende pouco sobre como solucionar o problema. No entanto, se a quantidade de épocas for muito alta, acarretará no *overfitting*, estado em que a rede encontra dificuldades para generalizar a solução do problema, de forma que, a partir de um determinado número de épocas, o erro da rede, ao invés de diminuir, aumenta considerável e continuamente, conforme a Figura 2.21.

Além disso, uma medida muito utilizada para avaliar problemas de classificação é a Acurácia, conforme a Equação 2.7

$$Acurácia = \frac{Número\ de\ Previsões\ Corretas}{Número\ Total\ de\ Previsões} \tag{2.7}$$

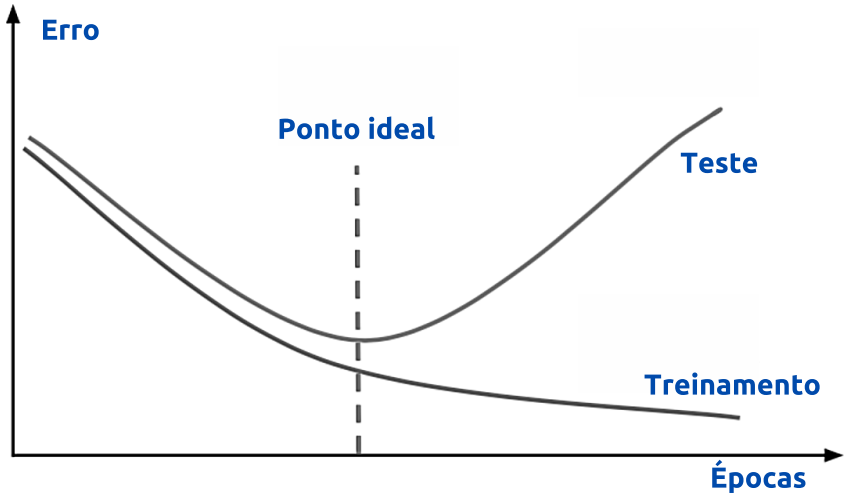


FIGURA 2.21 – Gráfico indicando a presença de *overfitting*, em relação ao número de épocas.

Com relação a problemas de classificação binária (sim ou não, por exemplo), é usada a **Matriz de Confusão** (POWERS, 2011), que é uma tabela que mostra as frequências de classificação para cada classe do modelo.

Utilizando-se dos valores da matriz de confusão, a Acurácia da RNA é definida pela seguinte Equação 2.8:

$$Acurácia = \frac{TN + TP}{TN + TP + FP + FN} = \frac{Previsões\ Corretas}{Total\ de\ Previsões} \tag{2.8}$$

O ***Recall*** é uma medida para analisar o quão bom o modelo é para se prever aquilo que esteja buscando e evitar perdas. Sendo assim, as amostras realmente negativas são descartadas. O *Recall* segue a Equação 2.9.

$$Recall = \frac{TP}{TP + FN} \tag{2.9}$$

2.5.8 Revisão das Arquiteturas de RNA

As características das RNA, seus parâmetros, forma de aprendizado e medidas de precisão formam o escopo básico de entendimento. Entretanto, como os modelos de RNA construídos neste estudo são compostos por alguns tipos de arquitetura de RNA específicas, convém detalhar as mesmas com maior quantidade de informações.

2.5.8.1 *Multilayer Perceptron*

Iniciando pelas Redes *Multilayer Perceptron* (MLP). As MLPs são compostas pela camada de entrada, saída e por mais de uma camada escondida. Possuem a capacidade de tratar com dados que não são linearmente separáveis e tem a sua arquitetura descrita na Figura 2.22.

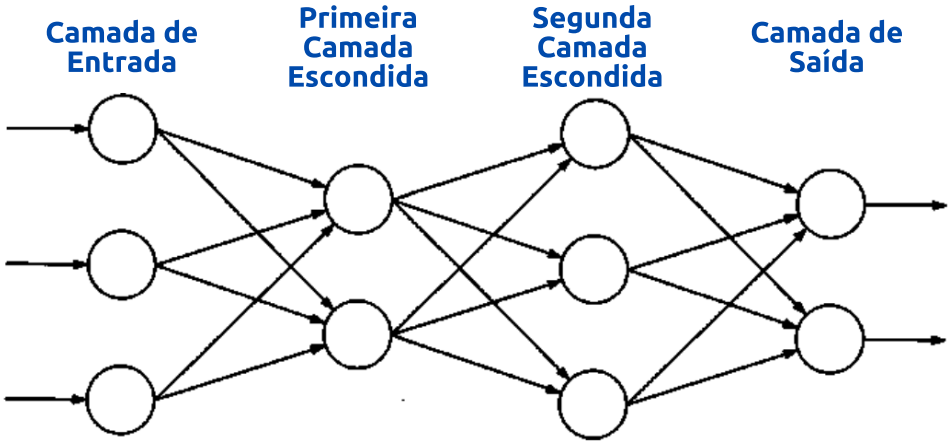


FIGURA 2.22 – Exemplo de Rede *Multilayer Perceptron*.

Um grande desafio para solucionar problemas que empregam redes MLP é encontrar a melhor configuração de número de nós e camadas que seja mais eficiente em questão de tempo, bem como encontrar valores o mais próximo da solução esperada. Nem sempre mais nós e camadas aumenta a eficiência da rede. Muitas vezes, este problema se resolve por meio de tentativas e erros.

Com relação ao número de nós, há diversas técnicas para resolver este problema. Entre elas, se destaca a Apodização (*Pruning*) (CERQUEIRA *et al.*, 2001), que consiste em cortar as conexões (ou pesos) dos neurônios que possuem pouca influência no erro, durante a etapa de treinamento. Dessa forma, reduz a complexidade da rede neural, melhorando sua capacidade de previsão. Além disso, contribui para diminuição da possibilidade de haver *overfitting*.

2.5.8.2 Rede Neural Convolucional

Em seguida, o próximo modelo é o da Rede Neural Convolucional (CNN). Este tipo de rede neural foi projetada para lidar com dados de imagem de forma eficiente. Também é reportado que seu uso é eficaz em tarefas como classificação de imagens, localização de objetos, leitura de legenda em Figuras, dentre outros. Adicionalmente, podem ser utilizadas para a resolução de outros problemas, como reconhecimento de atividade humana (YANG *et al.*, 2015) e classificação de séries temporais (LECUN *et al.*, 1995).

A CNN consiste em extrair automaticamente, a partir dos dados de entrada, as características que são úteis para a resolução do problema em questão, antes de ocorrer o processamento das informações por parte dos neurônios. Dessa forma, as CNNs tem se mostrado bastante eficientes, uma vez que os neurônios processam um número de dados menor do que o original. O pré processamento dos dados de entrada é composto por, basicamente, três ações (NIELSEN *et al.*, 2017):

- **Convolução** — Convolução é uma operação matemática para mesclar dois conjuntos de informações. No contexto de RNA, supondo que os dados de entrada sejam uma imagem onde cada pixel seja equivalente a um valor de entrada. É efetuada a multiplicação dela por um **filtro de convolução** (*kernel*), que indica alguma característica que se deseja encontrar na imagem, como detecção de borda, desfoque, relevo e etc. Ambos os elementos são matrizes de duas dimensões (ROSEBROCK, 2017). O resultado é denominado **Mapa de Características**, conforme Figura 2.23;

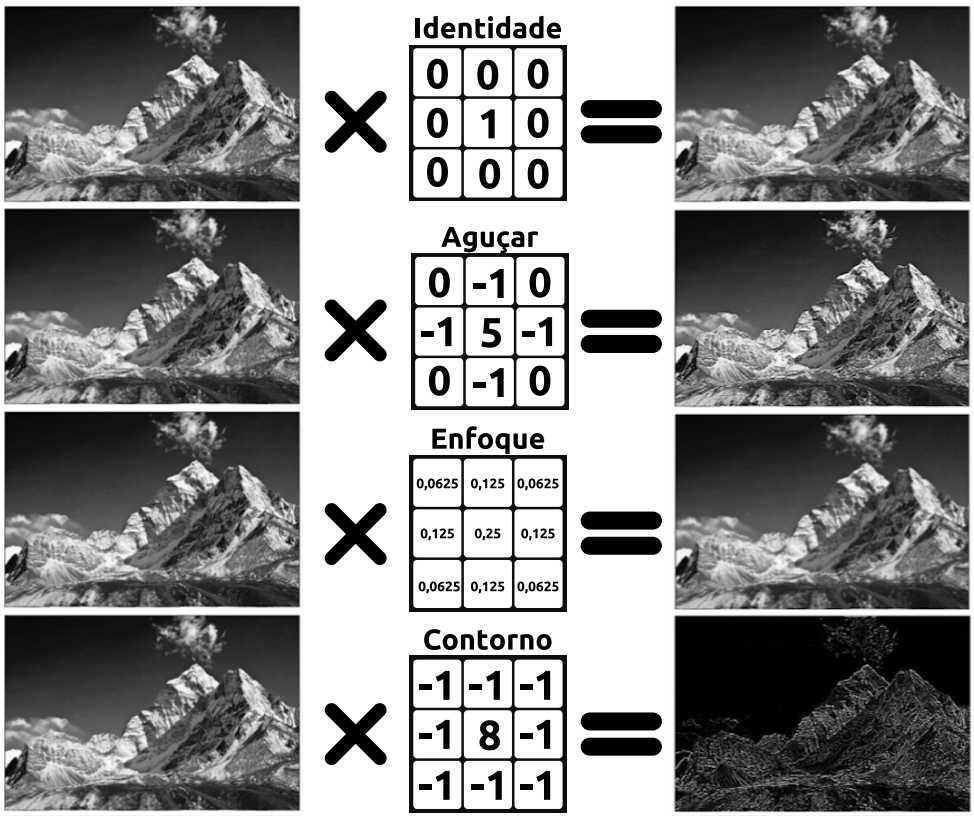


FIGURA 2.23 – Exemplo de Convolução de uma imagem.

- ***Pooling*** — Tem como objetivo, a partir do Mapa de Características, realçar suas principais características, de forma a diminuir ainda a mais as dimensões da matriz. A Figura 2.24 é um exemplo do uso do *max pooling*, que tem a finalidade de extrair os valores máximos de cada parte da matriz; e

- **Achatamento** — Consiste em transformar a matriz resultante do processo de *Pooling*

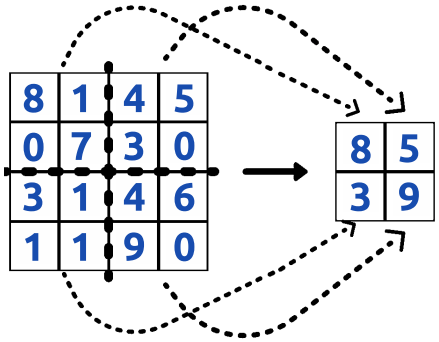


FIGURA 2.24 – Processo de Pooling

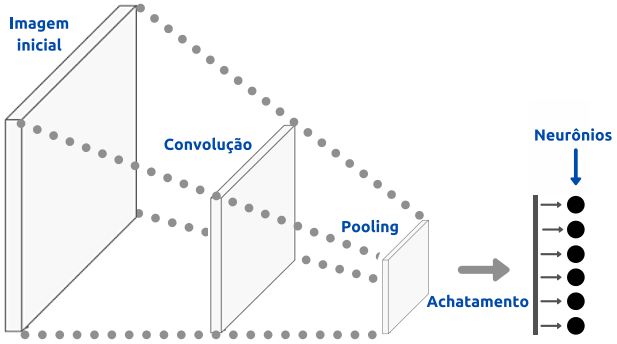


FIGURA 2.25 – Processo completo de Convolução de uma imagem

em somente uma dimensão. Em seguida, cada elemento será uma entrada em cada neurônio da RNA, acabando assim, a CNN. Poderá outras camadas após, como uma rede MLP, mas a CNN termina após o achatamento.

2.5.8.3 Long-Short Term Memory

Finalmente, como reportado na literatura, o modelo simples de RNN, algumas vezes, não consegue fazer boas previsões em sequências mais longas e complexas, tendo em vista que, por ocasião do processo de atualização dos pesos, o valor do gradiente diminui consideravelmente, ao ponto de não causar uma mudança significativa dos pesos. O nome desse problema se chama **Gradiente de Desaparecimento** (*Vanishing Gradient*) (KOSTADINOV, 2018).

Com a finalidade de resolver este problema, foi criado o modelo de Rede Neural LSTM *Long-Short Term Memory*, capaz de armazenar memória a longo prazo, sendo capaz de prover soluções que envolvam grandes dados sequenciais. Diferente de uma RNN simples, a unidade da rede LSTM não é um neurônio, mas sim, uma célula composta por quatro camadas, as quais são realimentadas ao longo do tempo (GOODFELLOW *et al.*, 2016) (Figura 2.26).

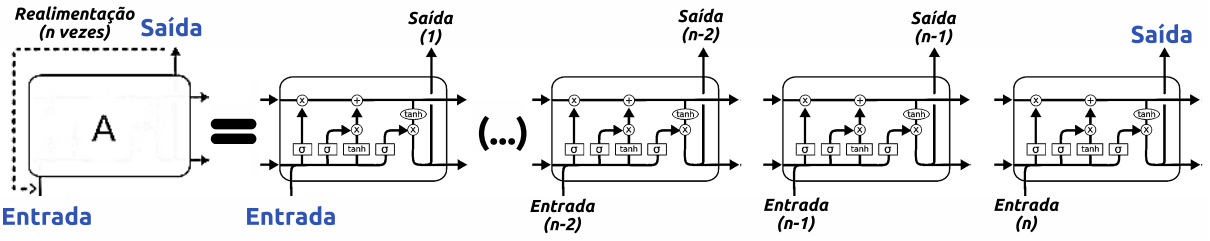


FIGURA 2.26 – Processo de realimentação de uma rede LSTM.

Cada célula contem o **Estado da Célula**, que é onde as informações são trafegadas ao longo do tempo. É representado pela linha horizontal na parte superior da Figura 2.27. O

Estado da Célula é constantemente atualizado por meio de sucessivas adições e remoções de informações, por ocasião de cada ciclo de realimentação.

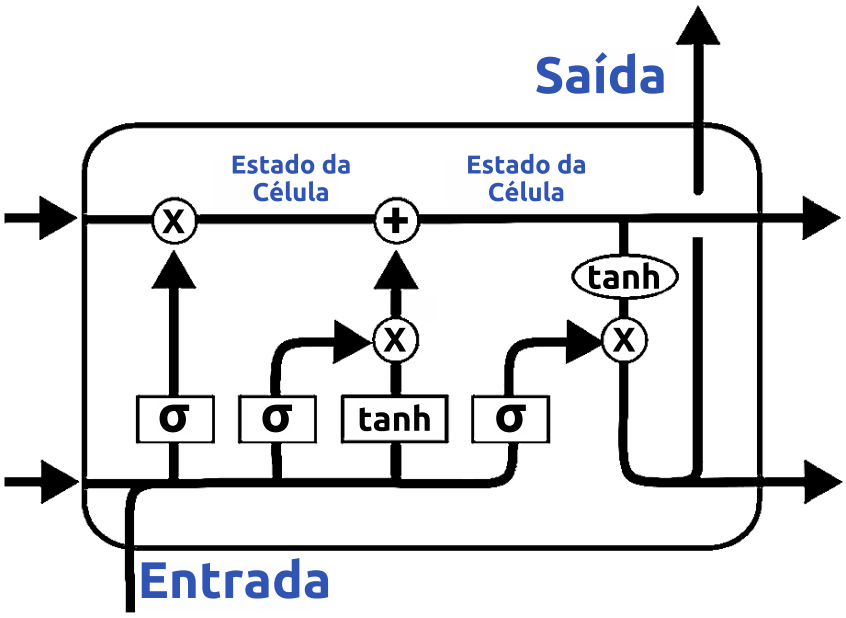


FIGURA 2.27 – Fluxo de Informações da Rede LSTM.

Tais informações são acrescentadas ou retiradas pelas **portas**, as quais são diferentes redes neurais que decidem quais informações serão persistidas no Estado da Célula. As portas deverão aprender quais informações são relevantes para guardar ou esquecer durante o treinamento da rede LSTM.

O processo de persistência de informações nas células da rede LSTM é efetuado pelos seguintes passos (Figura 2.28)[8] (GOODFELLOW *et al.*, 2016):

1. **Decidir o que será removido**. Pega a saída do tempo anterior e concatena com a entrada do tempo atual. Logo após, submete à função de ativação sigmóide, que decide qual dado será removido do Estado da Célula;

2. **Decidir o que será adicionado**. Em seguida, os mesmos dados são direcionados à função de ativação sigmóide e tangente hiperbólica, a fim de selecionar quais dados serão adicionados no Estado da Célula;

3. **Atualização do Estado da Célula**. As etapas anteriores decidiram o que apagar e o que armazenar, e agora essas etapas são executadas, mantendo o Estado da Célula atualizado; e

4. **Decidir as informações de saída**. Nesta última etapa, as informações passam por mais uma função de ativação sigmóide e uma tangente hiperbólica, a fim de selecionar quais os dados deverão ir para a próxima realimentação.

[8] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

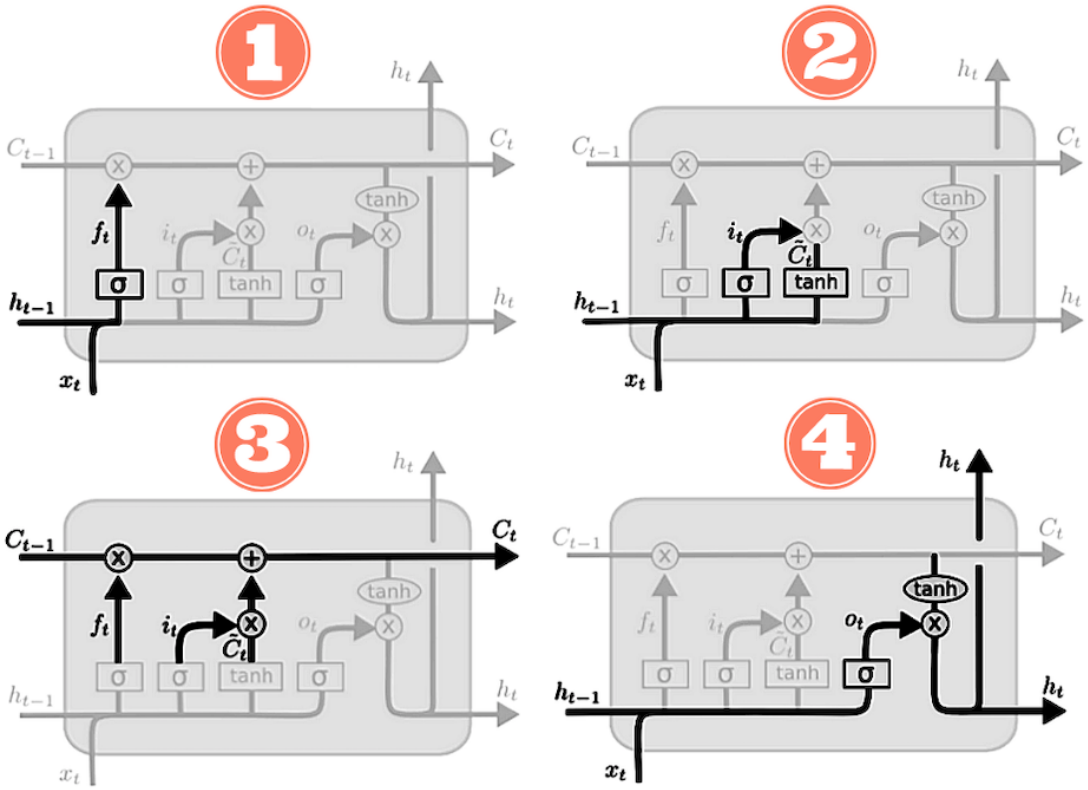


FIGURA 2.28 – Processo completo de persistência de informações numa Rede LSTM.

2.6 Aplicação da Fundamentação Teórica à Metodologia

Ao longo deste Capítulo, foram vistos os principais conceitos que servem de base para a construção do Método para Detecção de Fraudes em Criptomoedas. A Figura 2.29 mostra o Método, com seus cinco passos e as suas duas bases (Fundamentação Teórica e Trabalhos Relacionados), onde há a indicação de que a Fundamentação Teórica foi vista.

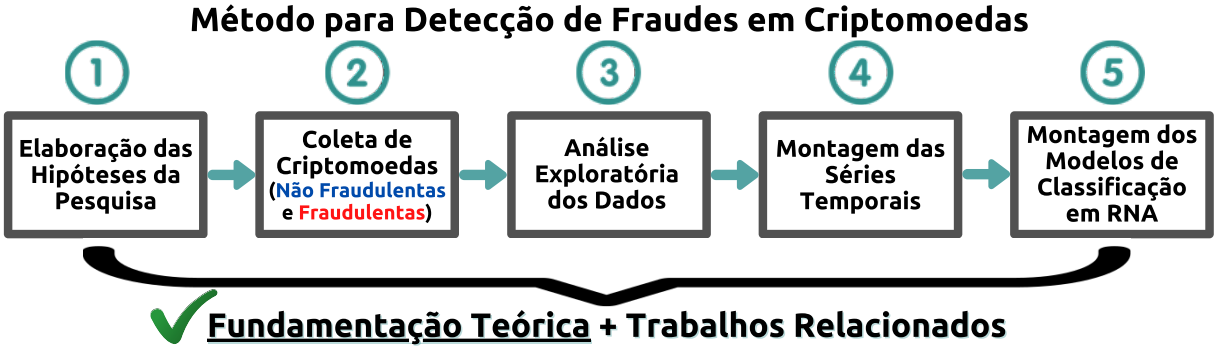


FIGURA 2.29 – Aplicação da Fundamentação Teórica ao Método Proposto.

A Fundamentação Teórica é dividida em cinco itens: Tecnologia *Blockchain*, Ethereum, Aspectos Econômicos de Criptomoedas, Séries Temporais e RNA. Cada item serviu de base de conhecimento para montar parte do Método, conforme a Tabela 2.1. O item Tecnologia *Blockchain* tem como objetivo auxiliar o entendimento de como funciona a

TABELA 2.1 – Correlação entre os itens da Fundamentação Teórica e os passos do Método de Detecção de Fraudes em Criptomoedas

Item da Fundamentação Teórica	**Passo do Método**
Tecnologia *Blockchain*	*Nenhum passo. É base para entendimento das características da rede Ethereum*
Ethereum	2) Coleta de Criptomoedas 3) Análise Exploratória dos Dados
Aspectos Econômicos de Criptomoedas	2) Coleta de Criptomoedas 3) Análise Exploratória dos Dados
Séries Temporais	4) Montagem das Séries Temporais
Redes Neurais Artificiais	5) Montagem dos Modelos de Classificação em RNA

rede Ethereum. Os itens Ethereum e Aspectos Econômicos de Criptomoedas são bases para o desenvolvimento dos passos: Coleta de Criptomoedas e Análise Exploratória dos Dados. Os itens Séries Temporais e Redes Neurais Artificiais, por sua vez, são bases para os passos, respectivamente, Montagem das Séries Temporais e Montagem dos Modelos de Classificação em RNA.

No entanto, para que se possa elaborar as hipóteses deste trabalho e fundamentá-las, bem como tornar a Metodologia o mais atual possível, é preciso utilizar-se dos conhecimentos recentes e que estejam no estado da arte desta área de pesquisa. Portanto, faz-se necessário o estudo dos Trabalhos Relacionados que se encontram no Capítulo 3.

3 Trabalhos Relacionados

Para a realização do levantamento bibliográfico, foi realizado uma revisão sistemática, classificação dos principais artigos e estudo dos mesmos. Neste Capítulo será explicado também o Método utilizado nesta pesquisa, bem como um resumo das fontes encontradas.

3.1 Método de Pesquisa

Inicialmente, destaca-se que as pesquisa de fontes de levantamento bibliográfico seguiram um padrão de busca criteriosa com base em buscadores online (IEEE, ACM, Springer, Elsevier, Portal CAPES e Google Scholar). Além disto, também seguiu-se e expandiu-se a coleta de referências bibliográficas, a partir dos artigos referenciados dos resultados dos buscadores online e referências avulsas. É possível descrever o processo de revisão sistemática aplicado em 3 fases:

1. Na primeira fase, foram procurados os assuntos *"cryptocurrency"*, *"fraud"* e *"initial coin offering"*, a fim de gerar uma base sólida de conhecimento sobre o assunto em questão. Desta forma, foram lidos diversos artigos e notícias sobre o assunto;

2. Após a primeira fase, foi observado que há artigos que propõem solução para fraudes em criptomoedas. Além disso, foi constatado que muitas soluções propostas se deram por meio de aprendizado de máquina. Sendo assim, foi realizada a segunda fase, com a busca pelos seguintes assuntos: *"cryptocurrency"*, *"fraud"* e *"neural network"*, com o objetivo de verificar como foi a aplicabilidade do conhecimento de redes neurais para a solução do problema. Foram observados os empregos de métodos de classificação baseados em redes neurais.

   No entanto, a maioria das fontes encontradas se atém aos aspectos sociais, como a existência (ou qualidade, caso exista) dos seguintes elementos: *site*, *"white paper"*, currículo no Github[1] dos membros da equipe de desenvolvimento, dentre outros. Tal solução apresenta pontos de falhas, pois os dados dependem do ponto de vista, muitas vezes subjetivos, dos avaliadores.

[1] https://github.com/

TABELA 3.1 – Resultados Encontrados em Revistas e Buscadores Científicos

Fonte	"cryptocurrency" "fraud" "initial coin offering"	"cryptocurrency" "fraud" "neural network"	"neural network" "time series classification" "ethereum"
IEEE	13	41	1
ACM	7	5	0
Springer	137	67	165
Elsevier	67	39	0
Portal CAPES	0	0	0
Google Scholar	1330	654	31

Por outro lado, um número bem reduzido de artigos se propôs a analisar o fluxo de transações na *blockchain*. Esta abordagem é mais segura, tendo em vista que os dados escritos na *blockchain* são imutáveis. Outro fator importante é que esta análise não depende de avaliação humana, mas somente dos dados contidos nas transações; e

3. Por último, foi pesquisado sobre *neural network*, *time series* e *classification*, a fim de consolidar conhecimento técnico sobre esta área. Foram encontrados alguns artigos e livros para este fim.

A Tabela 3.1 mostra um resumo quantitativo a cerca das pesquisas durante as três fases da Revisão. No total, foram selecionados 54 artigos. Seguem algumas observações qualitativas sobre esse processo:

- Os buscadores proprietários das revistas científicas não se mostraram eficazes, pois vários artigos foram encontrados pelo Google Scholar (e não por eles);
- Dentre os resultados encontrados, há poucos artigos científicos em português;
- Apesar de o pico de procura por ICOs ter sido em 2018, fraudes ainda é um assunto importante dentro da área de criptomoedas;
- A maioria dos artigos encontrados durante a fase 1 dizem respeito à área econômica, e não a de Computação; e
- Foram selecionados alguns livros, afim de prover uma base de conhecimento sólida sobre assuntos já consolidados, como Redes Neurais e Séries Temporais.

3.2 Principais Estudos Encontrados

Esta seção resume os principais estudos. Para cada estudo apresentado, traz-se um pequeno resumo onde se abordam as principais características. Os trabalhos relacionados estão organizados, primeiramente, os que focam mais no aspecto econômico e, em seguida, indo para os artigos mais próximos desta proposta. Em outras palavras, foram levantados desde *surveys* em atividades de ICO, até chegar a artigos mais voltados para detecção de fraudes e classificação em séries temporais.

3.2.1 Criptomoedas sob a Perspectiva da Escola Austríaca

Milne (MILNE, 2018) realizou um estudo a respeito das criptomoedas sob um ponto de vista da Escola Austríaca de Economia. O artigo fornece uma revisão histórica crítica do sistema monetário atual, uma descrição do ecossistema financeiro de criptomoedas e a sua relação com a Escola Austríaca no tocante à descentralização do Sistema Financeiro.

Adicionalmente, o estudo aborda que o crescimento da adesão das pessoas ao uso de criptomoedas se deu pelos seguintes fatores:

- Não há uma autoridade central, detentora de boa parte das reservas monetárias da moeda. Logo, um sistema centralizado teria um risco de ser interrompido, caso a sua autoridade central entre em falência;
- O dinheiro dos usuários ficam sob posse dos próprios usuários, e não por parte de terceiros, como um banco, por exemplo; e
- Evita que o sistema financeiro seja influenciado por decisões políticas.

3.2.2 Blockchain, Bitcoin e ICOs: uma revisão e guia de pesquisa

Romi Kher (KHER *et al.*, 2020) realizaram uma Revisão Sistemática, oferecendo um guia para pesquisas futuras em termos de fenômenos, teorias, metodologias e dados promissores, relativos à *blockchain* e ICOs. Além disso, inclui aspectos de regulação e segurança cibernética para combater fraudes em potencial.

O estudo apresenta uma explicação dos fundamentos da tecnologia *Blockchain* e suas aplicações como contratos inteligentes, criptomoedas, *tokens* e atividades de ICO. No total, foram revisados 152 artigos sobre o assunto, sobre diferentes áreas de conhecimento, como: Ciência da Computação, Economia, Empreendedorismo, Direito e Governança.

3.2.3 Por que os negócios estão indo para o "crypto"? Uma análise empírica de Oferta Inicial de Moedas

Adhami (ADHAMI *et al.*, 2018) relatam um estudo a respeito dos aspectos que contribuem para o sucesso de uma ICO em captar recursos para seu projeto. Foram analisadas 253 ICOs entre 2014 até agosto de 2017. É o artigo mais citado no Google Scholar sobre este tema.

Conforme sua análise, a disponibilidade de um *White Paper* bem elaborado não é uma característica relevante para potenciais colaboradores, ainda mais que eles não possuem qualquer validade jurídica. Por outro lado, a disponibilidade do código fonte, mesmo que parcialmente, é um fator de peso na escolha de potenciais colaboradores, uma vez que estes se mostram bastante familiarizados com tecnologia. Além disto, segundo o autor, a literatura sobre o assunto atesta que códigos abertos oferecem o potencial de se ter uma tecnologia mais flexível e aberta, sobrepujando, dessa forma, o argumento de que um código aberto seja mais suscetível à invasão.

Com relação a fase de Pré-ICO, embora o artigo não cite esse nome, há uma relação entre o sucesso de uma ICO e uma pré venda de *tokens* organizada. No entanto, não há relação entre o sucesso da ICO com mecanismos de bônus para a venda de *tokens*, embora o autor reconheça que se faz necessário um estudo mais aprofundado sobre quais esquemas de bônus de venda são relevantes para os investidores.

Sobre as possíveis funções dos *tokens*, o artigo elencou cinco:

- moeda;
- acesso a algum serviço;
- direitos de governança;
- direito de lucros; e
- direitos de contribuição com suporte.

Somente as funcionalidades relativas aos direitos de acesso a serviços e lucros apresentaram uma relação com o sucesso de uma ICO.

3.2.4 Avaliação de Oferta Inicial de Ativos Digitais: o estado da prática

Com o objetivo de fornecer maior segurança aos investidores, há disponíveis na internet vários sites que fornecem uma avaliação de diversas atividades de ICO. Hartmann

(HARTMANN *et al.*, 2018) fornecem um estudo de quais características de uma atividade de ICO os sites levam em consideração para avaliá-la. Foram elencados os seguintes pontos:

- Informações do Projeto (descrição, Clareza do *White Paper*, Canais de mídia Social, etc.);
- Informações da Equipe que compõe o Projeto;
- Informações do *tokens* a ser gerado (finalidade dos *tokens*, preço inicial e etc. . .);
- Informações sobre a atividade de ICO; e
- Informações técnicas como código fonte e *smart contracts*.

Além disso, o artigo apresenta uma classificação em termos da qualidade da avaliação provida por esses mesmos sites. Logo após, classificou os sites de acordo com dois critérios: transparência quanto ao processo de avaliação, ou seja, se deixa claro para o leitor quais características das atividades de ICO que foram levadas em consideração para efetuar a análise; e se o processo de avaliação é centralizado ou não (realizado pela própria equipe do site ou se é aberto ao público para que se possa avaliar).

3.2.5 Um estudo exploratório de contratos inteligentes na plataforma *Blockchain* da rede Ethereum

Gustavo A. Oliva (OLIVA *et al.*, 2020) efetuaram uma análise exploratória dos *smart contracts* presentes na rede Ethereum, com a finalidade de se ter uma ampla compreensão de todos os contratos lá implementados. Desta forma, foram utilizadas as seguintes ferramentas: Google BigQuery[2], Etherscan[3], State of the DApps[4] e CoinMarketCap[5].

O estudo obteve os seguintes resultados:

- Apenas 0,05% dos contratos inteligentes são responsáveis por 80% de todas as transações que envolvam contratos. Isto denota que as transações que circulam na Ethereum estão concentradas em uma parcela bem pequena de contratos. Além disto, 94,7% dos contratos receberam menos de 10 transações;
- A principal aplicação de contratos inteligentes é relativa ao gerenciamento e produção de *tokens*, o qual representa 41,3% de todas as transações. 72.9% de todos os contratos de alta atividade, contratos de *tokens*, possuem um *market cap* de, aproximadamente, US$12.7 bilhões; e

[2]https://console.cloud.google.com/bigquery
[3]https://etherscan.io/
[4]https://www.stateofthedapps.com/
[5]https://coinmarketcap.com/

- No que diz respeito à complexidade do código, foi observado que o código-fonte dos contratos classificados como de alta atividade é pequeno, com no máximo 211 instruções em 80% dos casos.

3.2.6 Detectando Esquemas Ponzi no Ethereum: Rumo a uma tecnologia *Blockchain* mais saudável

Chen (CHEN *et al.*, 2018), desenvolveram um método baseado em aprendizado de máquina, para detectar Esquemas Ponzi ao longo da rede Ethereum. Foram analisados 3071 contratos, baixados da plataforma Etherscan[6], nos quais duas características foram extraídas e serviram de base para a identificação de fraudes: as suas transações, incluindo as que tiveram erro, e seus respectivos códigos fonte.

Com relação às transações, foi observado que, em Esquemas Ponzi, grande parte dos usuários que recebem Ether são os primeiros titulares, ao passo que os demais investem recursos, mas não recebem (ou recebem pouco) Ether. Por outro lado, em um esquema não fraudulento, a distribuição entre investimento e recebimento de Ether é mais uniforme entre os usuários, ao longo do tempo.

Sob o ponto de vista de análise dos códigos fonte, um dos fatores mais preponderantes para a diferenciação entre contratos fraudulentos e não fraudulentos foi o GASLIMIT usado para a mineração do bloco.

Tais características foram utilizadas como dados de entrada em um classificador, que utiliza o algoritmo XGBoost. O índice de acurácia alcançou 81% (*recall*).

3.2.7 Explorando dados do *Blockchain* para detectar contratos de Esquemas Ponzi no Ethereum

Novamente, Chen (CHEN *et al.*, 2019), efetuaram um estudo sobre os contratos de Esquemas Ponzi. Dessa vez, foram incluídas variáveis como: número de participantes do contrato, investidores e recebedores, além do balanço de cada contrato.

O algoritmo usado pelo modelo de classificação foi o Random Forest. Um total de 3780 de códigos fonte de contratos e suas respectivas tabelas de transações foram analisados e foi obtido um resultado de 69% de acurácia (recall).

Ao final, o artigo estimou que há cerca de 500 contratos inteligentes de Esquemas Ponzi presentes na rede Ethereum.

[6]https://etherscan.io/

3.2.8 Dissecando Esquemas Ponzi no Ethereum: identificação, análise e impacto

Bartoletti (BARTOLETTI *et al.*, 2020) executaram um estudo sobre o comportamento de Esquemas Ponzi, dentro da *blockchain* Ethereum. Sendo assim, foram analisados 138 contratos relativos a criptomoedas desse tipo de atividade fraudulenta.

Sob o ponto de vista de análise dos códigos fonte dos contratos, foram detectados diversos esquemas, como: Pirâmide, Cascata, Corrente e Transferência. Além disto, foram detectados diversos *bugs* de código que, conforme o autor, boa parte é intencional.

Ao final do artigo, os autores resumem suas conclusões em três recomendações:

- Verifique as propagandas — Na maioria das vezes, Esquemas Ponzi apresentam altos retornos e omitem os riscos do negócio. Há sítios eletrônicos como o BadBitcoin[7] que contém uma lista negra de golpes baseados em criptomoedas;
- Analise o código do contrato — Muitos usuários possuem a mentalidade de que, uma vez que o código do contrato é transparente a todos, ele é confiável, o que não é verdade. A grande maioria dos contratos apresentam um código simples, de aproximadamente 100 linhas. Com a finalidade de mitigar esta vulnerabilidade, foram desenvolvidos programas que verificam o código do contrato, de forma automática; e
- Estude os *logs* das transações — Aproximadamente 60% dos Esquemas Ponzi apresentam tempo de vida menor que 1 dia. Sua análise exploratória de contratos de Esquemas Ponzi foi aplicada ao modelo classificador, desenvolvido por Chen (CHEN *et al.*, 2018), utilizando parâmetros como: número de pagamentos, saldo do contrato, proporção de investidores que recebeu pelo menos um pagamento, dentre outros parâmetros. Os resultados obtidos apresentam que as características dos contratos são mais discriminatórias do que as dos *logs* das transações.

3.2.9 A Anatomia dos Esquemas *Pump and Dump* em criptomoedas

Jiahua Xu (XU; LIVSHITS, 2019) apresentam um estudo detalhado de esquemas de *Pump and dumps* em criptomoeda. O artigo descreveu a anatomia de um ataque típico e depois foram investigados diversos aspectos de ataques reais a criptomoedas em oito meses em quatro *exchanges*. O estudo demonstra que este tipo de fraude move uma quantidade de dezenas de milhões de dólares em volumes falsos de negociação a cada mês.

[7] https://badbitcoin.org/

Ainda, revela que os organizadores deste esquema podem usar facilmente suas informações privilegiadas obter lucro em cima de outros usuários.

Foi criado um modelo de classificação utilizando o algoritmo Random Forest, a fim de identificar esquemas de *Pump and Dumps*. Os dados de entrada foram as próprias informações de mercado das respectivas criptomoedas, como volume de transações e preço. Por meio da observações em grupos de Telegram[8] e Discord[9], foram relatados os processos de captação de usuários por parte dos donos desses esquemas. Ao total, 412 criptomoedas foram analisadas.

O artigo mostrou que é possível, com modelos de aprendizado de máquina razoavelmente rudimentares, prever com precisão de aproximadamente 90% as criptomoedas alvo deste tipo de fraude. O estudo fornece uma prova de conceito para o uso de aprendizado de máquina para a detecção de fraudes no mercado de criptomoedas.

3.2.10 *Deep Learning* para Previsão de Séries Temporais

Jason Brownlee (BROWNLEE, 2018) desenvolveu um método para classificação de Séries Temporais, utilizando RNAs. Para isto, ele reuniu diversos artigos como (WANG *et al.*, 2019), (MURAD; PYUN, 2017), (ORDÓÑEZ; ROGGEN, 2016), (ZENG *et al.*, 2014) e (ANGUITA *et al.*, 2013), que abordam técnicas para Reconhecimento de Atividade Humana.

O livro descreve modelos de classificação que contém diferentes arquiteturas de RNA como: CNNs, MLPs, LSTMs e redes híbridas, como CNN-LSTM e ConvLSTM. Tais modelos utilizam uma ou mais Séries Temporais como dados de entrada. Sendo assim, é possível efetuar uma classificação de séries univariadas e multivariadas. Python foi a linguagem de desenvolvimento utilizada no livro.

Além disso, o autor realizou uma comparação entre o uso dessas arquiteturas de RNAs e algoritmos conhecidos em aprendizado de máquina, como Naive Bayes, KNN e Random Forest para a resolução de problemas que recebam Séries Temporais como entrada. Em geral, as RNAs obtiveram melhores resultados que os demais algoritmos.

3.3 Aplicação dos Trabalhos Relacionados ao Método

A Figura 3.1 mostra o Método, com seus cinco passos e as suas duas bases (Fundamentação Teórica e Trabalhos Relacionados), onde há a indicação de que a Fundamentação Teórica e os Trabalhos Relacionados foram vistos.

[8]https://web.telegram.org/
[9]https://discord.com/

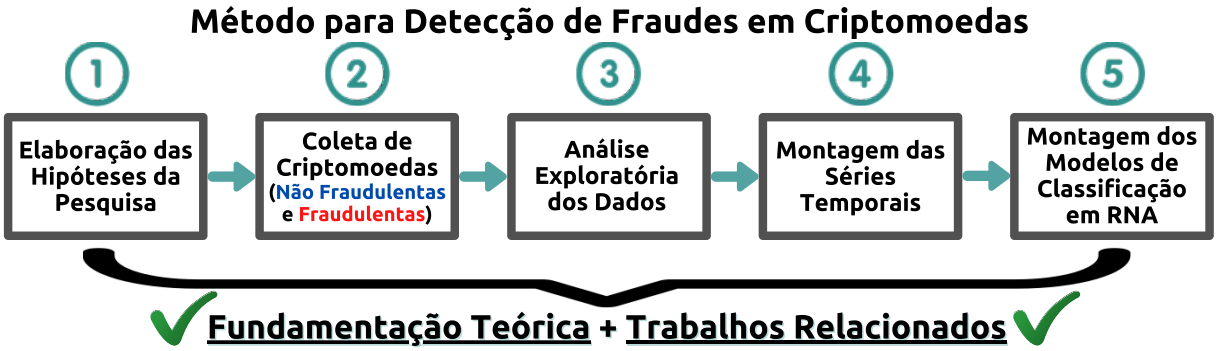


FIGURA 3.1 – Aplicação da Fundamentação Teórica e dos Trabalhos Relacionados ao Método Proposto.

TABELA 3.2 – Correlação entre os Trabalhos Relacionados e os passos do Método de Detecção de Fraudes em Criptomoedas

Trabalho Relacionado	**Passo do Método**
Criptomoedas sob a Perspectiva da Escola Austríaca	1) Elaboração das Hipóteses da Pesquisa
Blockchain, Bitcoin e ICOs: uma revisão e guia de pesquisa	2) Coleta de Criptomoedas
Por que os negócios estão indo para o “crypto”? Uma análise empírica de Oferta Inicial de Moedas	2) Coleta de Criptomoedas
Avaliação de Oferta Inicial de Ativos Digitais: o estado da prática	2) Coleta de Criptomoedas
Um estudo exploratório de contratos inteligentes na plataforma Blockchain da rede Ethereum	3) Análise Exploratória dos Dados
Detectando Esquemas Ponzi no Ethereum: Rumo a uma tecnologia Blockchain mais saudável	1) Elaboração das Hipóteses da Pesquisa 3) Análise Exploratória dos Dados
Explorando dados do Blockchain para detectar contratos de Esquemas Ponzi no Ethereum	3) Análise Exploratória dos Dados
Dissecando Esquemas Ponzi em Ethereum: identificação, análise e impacto	3) Análise Exploratória dos Dados
A Anatomia dos Esquemas Pump and Dump em criptomoedas	1) Elaboração das Hipóteses da Pesquisa 3) Análise Exploratória dos Dados
Deep Learning para Previsão de Séries Temporais	4) Montagem das Séries Temporais 5) Montagem dos Modelos de Classificação em RNA

Ao longo deste Capítulo, encontra-se uma revisão na literatura atinente ao tema deste trabalho, bem como os Trabalhos Relacionados relevantes, que contribuíram para o alcance do estado da arte do conhecimento a respeito de fraudes em criptomoedas. Os

Trabalhos Relacionados contém nove artigos somados a um livro. A correlação entre eles e os cinco passos do Método para Detecção de Fraudes encontra-se na Tabela 3.2.

Algumas observações importantes quanto a essa correlação são: 4 artigos sustentam a Elaboração das Hipóteses da Pesquisa; 3 artigos possuem grande contribuição para a Coleta de Criptomoedas, uma vez que descreve as características mais atuais sobre elas, tornando o Método mais próximo da realidade; e o livro sobre Previsão de Séries Temporais serve de base de conhecimento para a montagem das Séries Temporais e dos Modelos de RNA.

Tendo em vista que os conhecimentos da Fundamentação Teórica e dos Trabalhos Relacionados já estão abordados, torna-se possível estudar o Método proposto, que se encontra descrito no próximo Capítulo 4.

4 Método e Caracterização dos Dados

Como explicado no Capítulo 1 Introdução, o núcleo deste trabalho tem como objetivo detectar fraudes em criptomoedas, originadas de atividades de ICO na rede Ethereum, por meio do desenvolvimento de modelos de classificação baseados em RNA. Para atingir o objetivo, realizou-se um esforço para se obter os dados, tratá-los de maneira a tornar usáveis para modelos, realizar uma caracterização e também investigar relações entre variáveis dentro de diferentes conjuntos de hipóteses de fraudes.

Nesta Seção, serão abordados esses aspectos. Em detalhes, serão mostradas as principais características dos conjuntos de dados (*datasets*) utilizados. Sendo que estas características serão organizadas em torno do que denominou-se: hipóteses de características que indicam fraudes. Tais hipóteses foram inspiradas em intuição de fraudes semelhantes no mundo real.

Com base nessas hipóteses, foram organizados os dados em formato de séries temporais, extraídos a partir do fluxo público de transações de cada criptomoeda selecionada, onde a hipótese determina o parâmetro da série. E, por sua vez, criou-se manualmente a saída de cada série baseada em intensa inspeção, conforme descrito ao longo do texto. Finalmente, a saída do classificador é um valor *booleano* (fraudulenta ou não fraudulenta) usado no treinamento e testes.

Convém destacar que todos os códigos de montagem de séries temporais, modelos de RNAs, aquisição do banco de dados, bem como a tabela contando as informações das criptomoedas coletadas estão disponíveis no repositório público Github[1].

Este Capítulo contem cinco seções, onde cada uma representa um passo do Método proposto neste trabalho, a saber: Elaboração das Hipóteses da Pesquisa; Coleta de Criptomoedas; Análise Exploratória dos Dados; Montagem das Séries Temporais; e Montagem dos Modelos de Classificação em RNA.

[1] https://github.com/luizzmata/ICOFraudDetection

4.1 Elaboração das Hipóteses da Pesquisa

Para que seja possível compreender os fatores determinantes para a identificação de uma possível fraude, foram elaboradas 5 hipóteses sobre a importância de certos fatores, que podem ser traduzidos em parâmetros, e auxiliam na identificação de fraudes, conforme Figura 4.1. Tais hipóteses são baseadas na literatura dos Trabalhos Relacionados descritos no Capítulo 2. Inclusive, em geral, tais hipóteses representam um papel decisivo para orientar os usuários quanto ao investimento em uma determinada nova criptomoeda.

Cada hipótese é abreviada por meio de uma palavra chave representativa da mesma, para facilitar a sua identificação ao longo do texto.

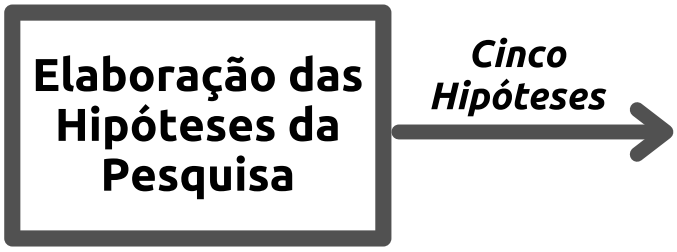


FIGURA 4.1 – Processo de Elaboração de Hipóteses da Pesquisa.

4.1.1 Inclusão de Novos Titulares (NEWHOLDER)

Hipótese 1: o fluxo de entrada de novos participantes, ao longo do tempo, é um fator relevante para detecção de fraudes.

Em esquemas fraudulentos, o momento em que é descoberta (ou revelada) a atividade irregular, acontece um movimento de redução da probabilidade de um novo investidor em comprá-la. Conforme a literatura base desta hipótese (XU; LIVSHITS, 2019), é comum que esquemas fraudulentos possuam um amplo fluxo de quantidade de transações no seu início, ao passo que, após a fraude ser descoberta, a quantidade de transações sofrer uma brusca redução, mantendo-se relativamente constante, ao longo do tempo.

4.1.2 Fluxo de Transações de Usuários Recém Criados (NEWUSER)

Hipótese 2: O fluxo de transações de usuários recém-criados na rede Ethereum, ao longo do tempo, é um fator relevante para detecção de fraudes.

Em esquemas de *Pump and Dumps*, de acordo com (XU; LIVSHITS, 2019), há um volume alto de transações feitas por usuários recém-criados, aqueles usuários cuja primeira transação será nessa série de ICO da rede *Ethereum*. O objetivo é ludibriar o público, dando a impressão de que a criptomoeda teve alta adesão no mercado, porque múltiplos

usuários estão comprando a mesma. Por outro lado, em esquemas não fraudulentos, o fluxo de transações de novos usuários apresenta uma distribuição mais uniforme ao longo do tempo, e não em rajada.

4.1.3 Maior Detentor de Títulos (BIGHOLDER)

Hipótese 3: O valor acumulado da parte que o maior detentor de ativos da criptomoeda possui, ao longo do tempo, é um fator relevante para detecção de fraudes.

Se um usuário for detentor de uma grande quantidade relativa de ativos, há uma alta possibilidade de o mesmo ser o dono do empreendimento por trás da criptomoeda e, portanto, quem tomaria todas as decisões de maneira centralizada. Sendo assim, há o risco de ele vender os ativos e abandonar o negócio quando o valor subir. No entanto, em esquemas não fraudulentos, a distribuição de ativos entre os usuários tende a ser descentralizada pois, conforme (MILNE, 2018), a descentralização das reservas monetárias é um atrativo para que as pessoas venham a aderir ao mercado de criptomoedas.

4.1.4 Taxas das Transações (GAS/GASLIMIT)

Hipótese 4: Os atributos das taxas de transações do Ethereum chamadas de GAS e GASLIMIT, ao longo do tempo, são fatores relevantes para detecção de fraudes.

Em geral, usuários maliciosos geram um volume maior de transações em relação aos não maliciosos. Por sua vez, cada transação possui um custo de execução e, consequentemente, quanto mais transações, maior o custo. Como o custo é maior, a intuição é que há uma maior preocupação por parte dos usuários maliciosos em evitar perdas. Ressalta-se que o GASLIMIT é um valor configurado pelos mineradores e, conforme apresentado nos Trabalhos Relacionados (CHEN *et al.*, 2018), possui um peso relevante para identificação de Esquemas Ponzi, bem como o atributo GAS, que é a própria taxa da transação.

4.1.5 Tempo após a Data de Entrada no Mercado (MARKETDATE)

Hipótese 5: A janela de tempo entre a data da entrada da criptomoeda no mercado e a data da sua análise é um fator preponderante para detecção de fraudes.

De acordo com os trabalhos relacionados das hipóteses anteriores, todos os esquemas fraudulentos apresentam uma redução brusca na quantidade de transações, no momento

em que a fraude é descoberta. Portanto, ao ser analisado o fluxo de transações de uma criptomoeda, desde a data de sua entrada no mercado, quanto maior o espaço de tempo em que as transações continuem acontecendo, maior será a probabilidade de detectar fraudes. Por outro lado, quanto menor a janela, mais difícil será de detectar atividades irregulares. Este parâmetro é determinante para realizar estudos em diferentes janelas de tempo para todas as séries estudadas, como, por exemplo, 20 dias ou mais.

4.1.6 Resumo das Hipóteses

Ao todo, cinco hipóteses norteiam este trabalho: NEWHOLDER; NEWUSER; BIGHOLDER; GAS/GASLIMIT; e MARKETDATE. A correlação entre elas e os Trabalhos Relacionados está descrita na Tabela 4.1.

TABELA 4.1 – Correlação entre as hipóteses e os Trabalhos Relacionados

Hipótese	**Trabalho Relacionado**
NEWHOLDER	A Anatomia dos Esquemas Pump and Dump em criptomoedas
NEWUSER	A Anatomia dos Esquemas Pump and Dump em criptomoedas
BIGHOLDER	Criptomoedas sob a Perspectiva da Escola Austríaca
GAS/GASLIMIT	Detectando Esquemas Ponzi no Ethereum: Rumo a uma tecnologia Blockchain mais saudável
MARKETDATE	*Todos os anteriores*

4.2 Coleta de Criptomoedas

Para a realização deste estudo, optou-se por capturar dados de transações sobre os processos de ICO, usando dados da blockchain principal da Ethereum. A obtenção de dados foi realizada entre 01/07/2020 e 01/08/2020, o que limita análises em novas criptomoedas. Além disto, é notório que existem milhares delas criadas todo mês. Desta forma, realizar a coleta das criptomoedas mais relevantes é uma atividade não-trivial. Desta forma, foram criados processos para a realização desta atividade, conforme Figura 4.2.

4.2.1 Pesquisa por Criptomoedas

Todas as criptomoedas que fazem parte deste estudo seguiram alguns critérios de relevância, conforme descrito na lista abaixo:

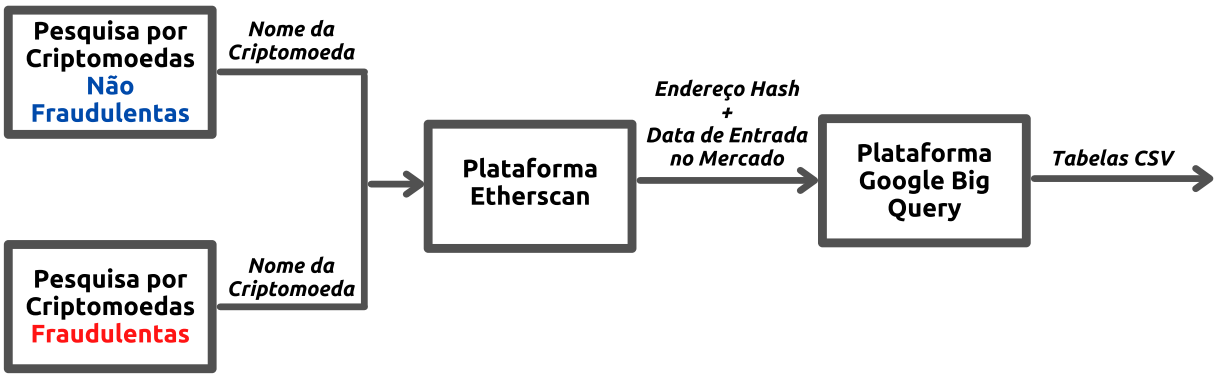


FIGURA 4.2 – Processo de Coleta de Criptomoedas.

- Ser um *token* na rede Ethereum, devido ao escopo da pesquisa;
- Ter, no mínimo, 6 meses de tempo de entrada no mercado;
- Possuir, no mínimo, 1500 transações, a fim de filtrar as criptomoedas que possuem um número de transações muito baixo. Esta quantidade de transações é necessária para se ter um conjunto suficiente de fluxo de transações para elaborar uma série temporal; e
- Devido às inconsistências nas datas de entrada no mercado, providas por diversas plataformas de notícias sobre atividades de ICO, utilizou-se, ao invés da data de entrada anunciada no mercado, a data que indica quando o número de transações atingiu o número mínimo de 400 transações, como indicador do primeiro dia. A plataforma Etherscan[2] foi utilizada para este fim.

Com base nos critérios acima, foram selecionadas 238 conjuntos de dados de ICOs de criptomoedas. O tamanho de cada conjunto de dados (*dataset*) de transações, em formato CSV, variou desde 450KB até 450MB, aproximadamente. Com base nestes *datasets* foram estabelecidos critérios, baseados em literatura e forte validação manual para especificar quais das criptomoedas podem ser consideradas, com segurança, como fraudulentas e não fraudulentas.

4.2.1.1 Seleção, Filtragem e Validação Manual de criptomoedas não fraudulentas

Para a utilização de modelos de aprendizado de máquina em séries temporais, é preciso estabelecer a variável alvo (*target*) de classificação, de modo preciso. Para tanto, foram utilizado os seguintes critérios de validação para a seleção de criptomoedas não fraudulentas:

[2]https://etherscan.io

1. Estar classificada entre as quatrocentas primeiras criptomoedas, de acordo com a plataforma Coingecko[3], em ordem decrescente de *Market Cap*. A escolha desta plataforma de ranqueamento se deu pelo fato de estar melhor pontuada na plataforma Alexa Rank[4]. A Alexa Rank é a plataforma que registra os sites mais populares, e portanto, mais confiáveis, da Internet. O *Market Cap* é um indicador que a criptomoeda vem sendo consolidada no mercado e, portanto, apresenta menor risco de fraudes, uma vez que cada titular fiscaliza seus recursos, e seu valor de mercado aumenta com o tempo. Em outras palavras, os usuários procuram investir em criptomoedas que não sejam fraudulentas, e portanto, seu valor se consolida com o tempo;

2. Estar na lista de atividades de ICO disponível na plataforma Icodrops[5]. Esta plataforma também encontra-se bem pontuada na plataforma Alexa e traz informações para investidores. Esta medida filtrou somente as criptomoedas que vieram de atividades de ICO; e

3. Finalmente, não ter qualquer notícia sobre fraudes, a partir da pesquisa na plataforma Google utilizando a seguinte busca: "ICO" + "scam" + "nome da criptomoeda". O sub-conjunto de criptomoedas que passaram nos critérios anteriores, mas não passaram neste critério, foi descartado.

Com base nesses estritos critérios de seleção obtivemos e validamos 102 ICOs que consideramos como não-fraudulentas. Ou seja, 42% do total. Isso é um indicador que o mercado de ICOs é muito suscetível para fraudadores.

4.2.1.2 Seleção, Filtragem e Validação Manual de criptomoedas fraudulentas

Novamente, foi utilizado um conjunto de critérios rigorosos para filtragem e marcação precisa das criptomoedas fraudulentas, a partir do conjunto de dados inicial. São eles:

1. Estar presente na lista de atividades de ICO fraudulentas dos sítios: coincurb[6] ou deadcoins[7] ou isthiscoinascam[8] ou etherscamdb[9] ou coinsopsy[10]. Tais *websites* foram resultados da pesquisa realizada na Plataforma Google, utilizando a seguinte busca: "ico"+ "scam"+ "list". Foi utilizado o critério inicial de aceitar primeiramente como verdadeiro, a maior quantidade de indicações de fraude possível;

[3]https://www.coingecko.com
[4]https://www.alexa.com/siteinfo
[5]https://icodrops.com/ico-stats/
[6]https://www.coincurb.com/deadcoin/
[7]https://deadcoins.com/
[8]https://isthiscoinascam.com/
[9]http://etherscamdb.info/scams
[10]https://www.coinopsy.com/dead-coins/

2. Ter ao menos um tópico sobre fraudes no fórum bitcointalk[11], fórum melhor classificado no Alexa Rank, a fim de aumentar a confiabilidade do item anterior; e

3. Não ter rede social atualizada ou não ter *websites* atividade de ICO disponível com detalhes da mesma. Tais medidas estão de acordo com as informações descritas nos Trabalhos Relacionados. A intuição geral é que as atividades de ICO fraudulentas não possuem *website* disponível, ou redes sociais desatualizadas (ou inexistentes). Foram obtidas estas informações gerais como páginas *web* e redes sociais através da plataforma anyblock[12] que vincula este tipo de informação.

4.2.1.3 Criptomoedas que não se encaixaram nos critérios deste estudo

Cerca de 80% das criptomoedas previamente analisadas foram descartadas por não se encaixarem nos critérios acima descritos. Tal acontecimento se deve ao fato de que, conforme o estudo (OLIVA *et al.*, 2020), a maioria dos contratos inteligentes em Ethereum possuem, no máximo 10 transações, o que inviabiliza a classificação de criptomoedas, sob a perspectiva de análise de séries temporais. Além disso, uma parcela pequena das criptomoedas não fraudulentas foi descartada, por não ter sido originada de atividades de ICO. Isto se deve ao fato de serem moedas *stable coins*, as quais são valoradas de acordo com a moeda. Por exemplo, o Tether é uma criptomoeda *stable coin* que vale sempre 1 dólar.

4.2.2 Aquisição dos bancos de dados de transações

As bases de dados são originárias de dados da rede Ethereum, conforme é possível verificar no *screenshot* da Figura 4.3. A tela é do aplicativo Etherscan, que permite uma visualização na rede Ethereum em formato *web*. Nela, há os detalhes de uma transação, contendo os endereços origem e destino Ethereum, do usuário que quer comprar *tokens* para o destino do contrato inteligente (*smartcontract*). Por meio da Plataforma Etherscan, foi possível, a partir dos nomes das criptomoedas, adquirir seus endereços *hash* e a Data de Entrada no Mercado.

Com as criptomoedas selecionadas, categorizadas conforme critérios descritos acima e com seus respectivos endereços *hash* e datas de entrada no mercado, foi realizado o processamento dos dados. Algo bastante conveniente nesse processamento foi a possibilidade do uso de *queries* específicas diretamente na plataforma Google Big Query[13]. A Google coleta muita informação, em especial, todas as transações na rede Ethereum estão

[11] https://bitcointalk.org/
[12] https://explorer.anyblock.tools/
[13] https://console.cloud.google.com/bigquery

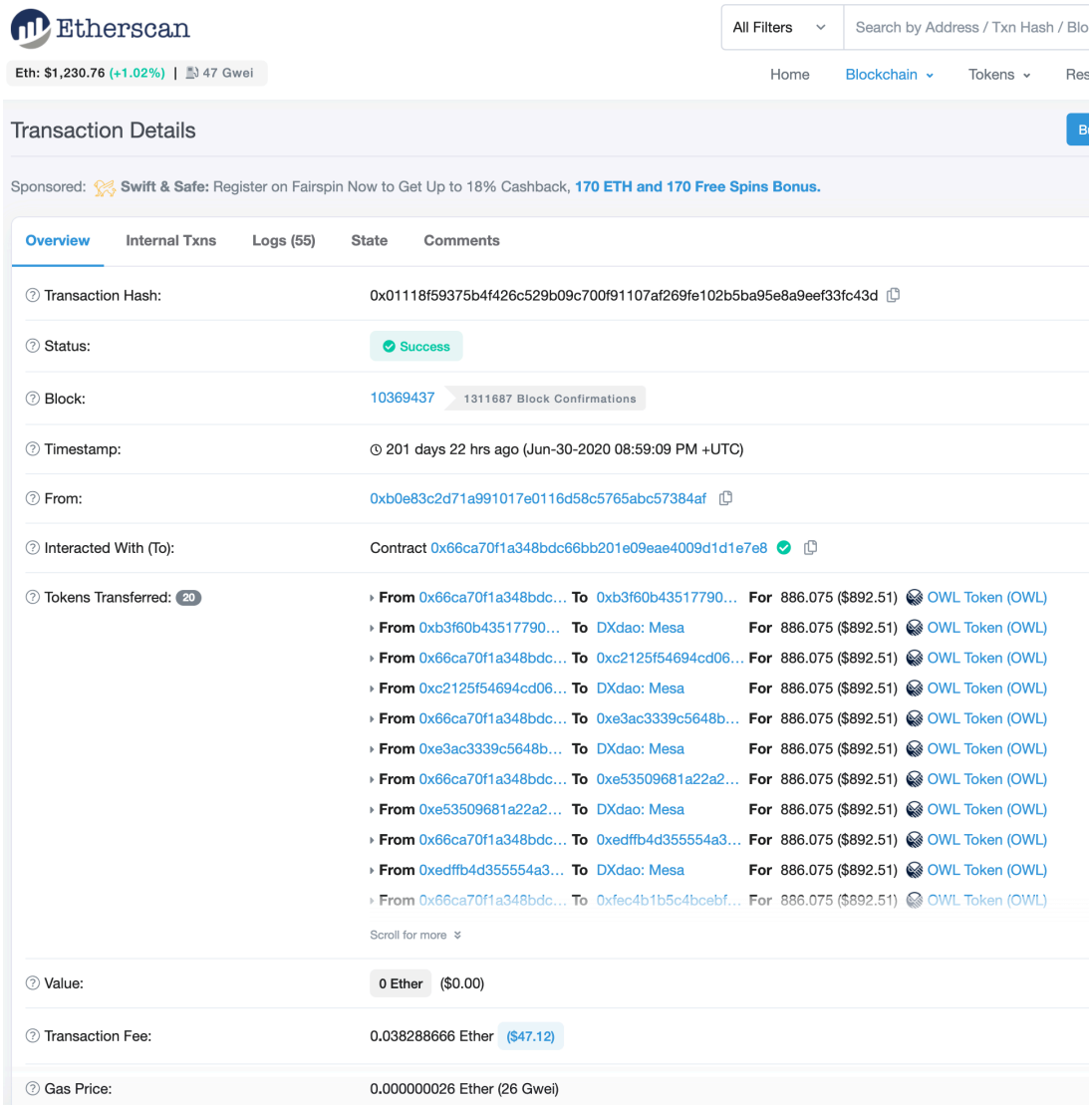


FIGURA 4.3 – Transações de criptomoedas descritas na Plataforma Etherscan.

armazenadas na Plataforma, o que permite uma fácil seleção e manipulação dos dados por meio de buscas específicas. Um exemplo de busca no Google Big Query está apresentada a seguir, onde "X" é o endereço *hash* da criptomoeda e "Y" é a data de entrada no mercado, acrescida de 180 dias (as tabelas CSV a serem trabalhadas terão o espaço de 180 dias de análise):

```
SELECT A.BLOCK_TIMESTAMP, A.FROM_ADDRESS, A.TO_ADDRESS, A.VALUE, A.
    TRANSACTION_HASH, B.NONCE, B.FROM_ADDRESS as FROM_ADDRESS_BLOCKCHAIN,
     B.TO_ADDRESS as TO_ADDRESS_BLOCKCHAIN, B.GAS, B.RECEIPT_GAS_USED
    FROM bigquery-public-data.crypto_ethereum.token_transfers as A INNER
    JOIN bigquery-public-data.crypto_ethereum.transactions as B ON A.
    transaction_hash = B.hash WHERE A.TOKEN_ADDRESS="X" AND A.
    BLOCK_TIMESTAMP<"Y 00:00:00 UTC"
```

Com base nesses dados obtidos pela plataforma Big Query, realizou-se a filtragem dos campos relevantes até a formação de um esquema único de armazenamento para este

trabalho. Este esquema de armazenamento é detalhado na Tabela 4.2, observando-se que somente foram processados os dados obtidos referentes aos seis primeiros meses após a data de lançamento no mercado. Os campos escolhidos para processamento incluem o *timestamp* que é insumo obrigatório para as séries temporais, em formato padronizado de tempo.

Os demais campos capturam as várias intuições descritas anteriormente, Endereços ethereum dos usuário e do *smartcontract* são os primeiros campos. Uma observação sobre o formato do endereçamento é que este é um valor em hexadecimal criado a partir do algoritmo padrão Ethereum *Keccak-256*. Em seguida, são apresentados os valores de transações e taxas (GAS) e também os endereços dos *tokens*. Toda essa organização forma a base para as séries temporais.

O valor de NONCE também é capturado, bem como os dados de GASLIMIT que são utilizados para detectar duplicatas, e também a influência do GASLIMIT conforme discutido anteriormente. Estes valores são em termos de inteiros e somente valores de transações em Ethereum (que podem ser convertidos para dólares ou reais) representam um valor fracionário de alta precisão.

Campo	**Descrição**	**Tipo de Dado**
Timestamp	Data e Hora	ISO 8601
From Address	Endereço *hash* do usuário origem da transação	Keccak-256
To Address	Endereço *hash* do usuário destinatário da transação	Keccak-256
Transaction Value	Valor da transação, na criptomoeda	integer
Transaction Address	Endereço *hash* da transação	Keccak-256
Transaction Nonce	Contador de Nonce	integer
Tkn Src Address	Endereço *hash* do usuário origem da transação de token	Keccak-256
Tkn Dst Address	Endereço *hash* do usuário destinatário de token	Keccak-256
GASLIMIT	Valor limite de GAS selecionado pelo minerador do bloco	integer
GAS	Taxa da transação usado pelo minerador do bloco	integer

TABELA 4.2 – Campos das Bases de Dados das transações de criptomoedas.

4.3 Análise Exploratória dos Dados

Nesta Seção, encontra-se a Análise Exploratória dos Dados (EDA), a fim de aprimorar as hipóteses elaboradas à luz dos dados provenientes das tabelas de transações. Neste intuito, foram estudados vários aspectos dos dados, por exemplo, dia da primeira transação, maiores detentores de títulos, volume total de transações, GAS e GASLIMIT por transação. Na maioria das vezes, as distribuições tem algumas diferenças importantes que poderão ser capturadas pelos modelos de redes neurais, a serem apresentados no próximo capítulo. Outras distribuições podem ser usadas no futuro como um segundo critério, que não é baseado em série temporal, para detecção de fraudes. Iniciando-se por uma análise

de datas, observando-se estar-se lidando com os 238 *datasets* neste estudo, devidamente classificados, segundo critérios validados de fraude. A Análise Exploratória de Dados foi realizada de acordo com o processo da Figura 4.4, onde as entradas são as tabelas CSV, sob o direcionamento de se investigar os dados relativos às hipóteses e a saída, o conhecimento para a montagem das séries temporais.

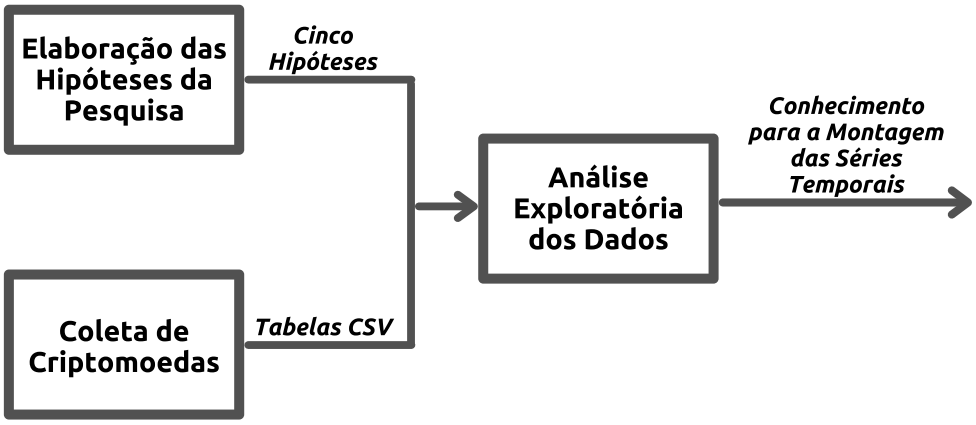


FIGURA 4.4 – Processo de Análise Exploratória de Dados.

4.3.1 Diferença entre a Data da Primeira Transação e Entrada no Mercado

Embora este aspecto não possa ser capturado em uma série temporal, é esperado que criptomoedas fraudulentas tenham uma preparação da fraude mais elaborada até a sua largada no mercado. Deste modo, foi realizado um levantamento das diferenças entre as datas da primeira transação na rede Ethereum e entrada no mercado para cada criptomoeda. Em geral, a primeira transação é basicamente uma inicialização do *smartcontract* com algum valor total de *tokens*.

De qualquer forma, como é notado nas Figuras 4.5a e 4.5b, tem-se o boxplot da distribuição dos pontos de diferença de datas. No caso das criptomoedas não fraudulentas, pode-se notar que a diferença entre a entrada no mercado e a primeira transação é bem pequena, com mediana inferior a 10 dias e com terceiro quartil inferior a 40 dias.

No caso das criptomoedas fraudulentas, elas apresentam maior diferença de dias entre essas datas, com mediana superior a 30 dias e uma variabilidade muito maior. Em ambas as classificações de ICOs, os valores de *outliers* são bastante extremos, superiores a 200 dias, indicando que o público alvo, as vezes, não se interessa prontamente por pequenas atividades de ICO.

(a) **Criptomoedas Não Fraudulentas** (b) **Criptomoedas Fraudulentas**

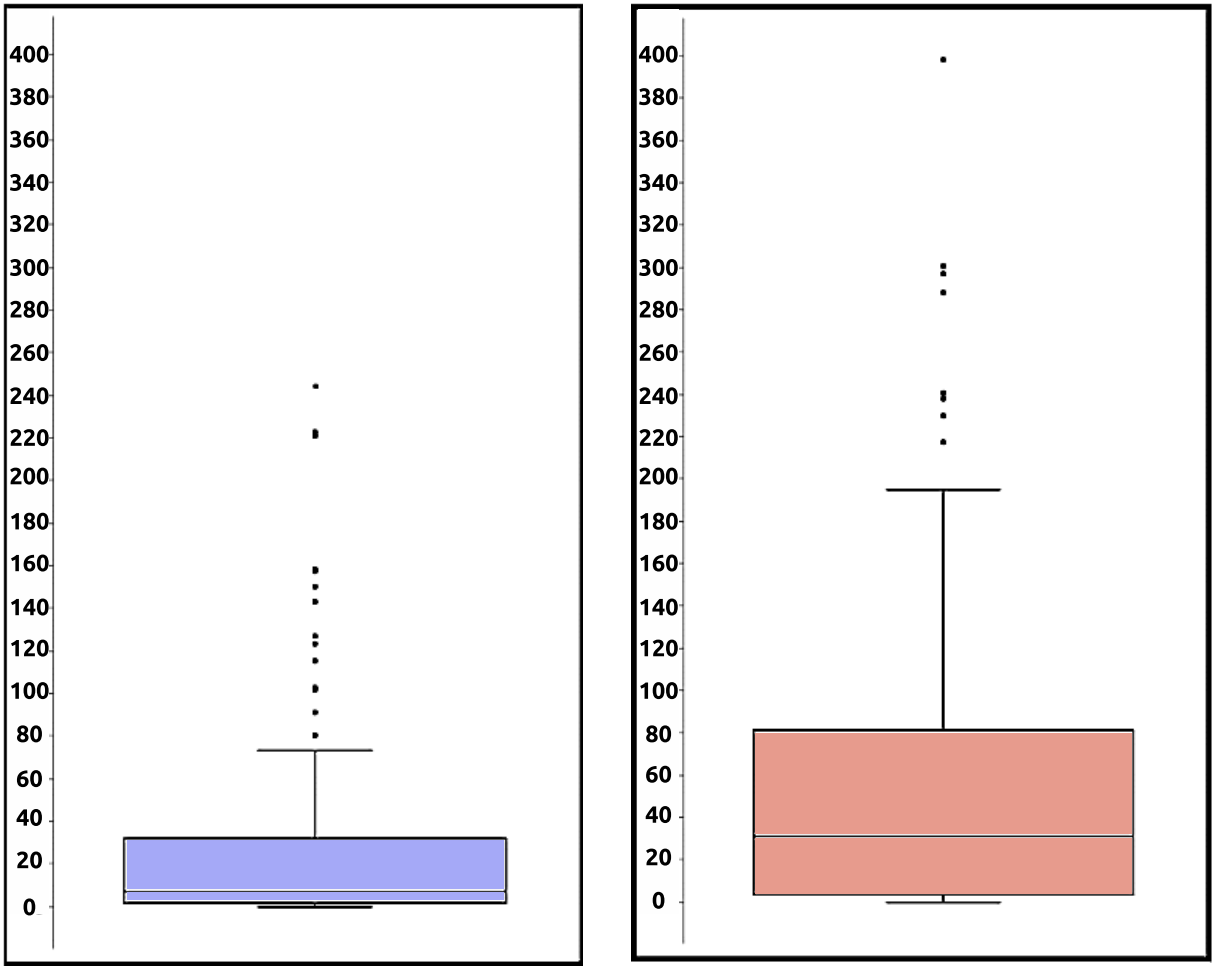


FIGURA 4.5 – Diferença em dias entre a data da primeira transação e a entrada no mercado para cada criptomoeda

4.3.2 Tipo de Conta do Maior Detentor de Títulos

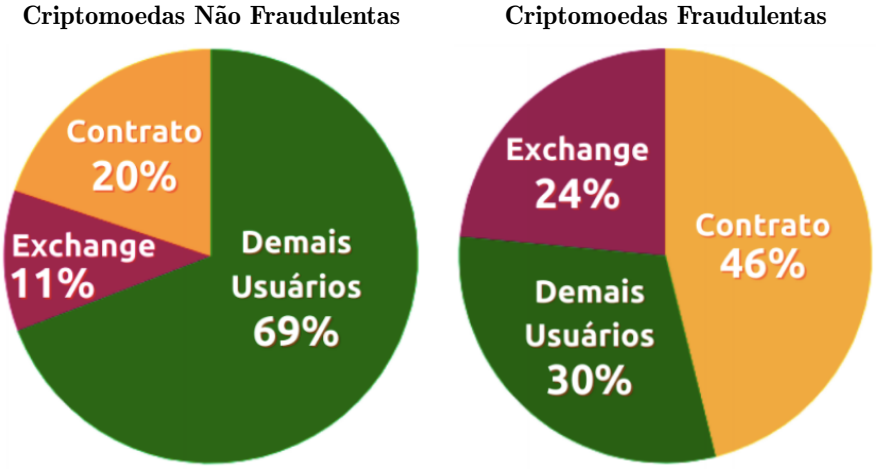


FIGURA 4.6 – Tipo de conta do maior detentor de títulos

Outro ponto a ser levado em consideração é entender qual o papel do tipo de conta (endereço *hash*) que configura os maiores detentores de títulos. O maior detentor de título, no escopo desta pesquisa, é a conta que detém a maior quantidade de ativos de uma determinada criptomoeda. Ela pode ser de três tipos: contrato (incluindo casas de câmbio descentralizadas), casa de câmbio centralizada ou demais usuários (os que não se encaixam nos dois anteriores). Analisando-se os *datasets* nota-se que muitas transações são feitas para a conta do *smartcontract*, o ponto intermediário principal entre os usuários que querem adquirir *tokens* durante a atividade de ICO. Outros endereços, o próprio Etherscan atribui a *exchanges* ou casas de câmbio para as criptomoedas, que, em geral, trabalham com volumes grandes de transações, a fim de realizar uma oferta de compra e venda, mediante especulação. E os demais usuários seriam contas distintas com endereços únicos comprando os títulos.

Para apresentar esses dados, fez-se uso de gráfico em pizza, conforme descritos nas Figuras 4.6a e 4.6b. Para as criptomoedas fraudulentas, é mais comum que os maiores detentores de títulos sejam contas que não sejam nem contrato e nem pertencentes a uma casa de câmbio (69% de usuários distintos). Por outro lado, para as criptomoedas não fraudulentas, existe a maior chance de sucesso de elas serem publicadas em Casas de Câmbio para negociação, o que impulsiona o valor das mesmas e aumenta a sua chance de se estabelecer no mercado. Nesses casos, somando-se os detentores de títulos do próprio *smartcontract* e da *exchange*, o total é o contrário das criptomoedas fraudulentas, com 70% desses tipos de endereços.

4.3.3 Histograma da Dispersão de Tempo entre as Transações

As transações e seus respectivos tempos de realização são a base das séries temporais. Portanto, é importante estudar como é a distribuição da disperção dos tempos entre transações, ou seja, em quais gráficos de transações de criptomoedas as dispersões são mais frequentes, onde os usuários interessados se apresentam com maior velocidade. Ambas as categorias apresentaram uma frequência enorme de dispersão entre transações (tempo entre duas transações consecutivas) nos blocos de tempo de 1 dia até 10 dias, conforme histograma da Figura 4.7. No eixo-x encontra-se a quantidade em dias entre transações e, no eixo-y, a contagem da frequência dessas. Sendo assim, foi utilizado o valor de 1 dia como referência para computar as janelas de dados das séries temporais.

4.3.4 Quantidade de transações totais

Outro aspecto que chama a atenção nas criptomoedas classificadas é que elas exibem diferenças em termos de quantidades totais de transações, devido ao fato de as criptomoe-

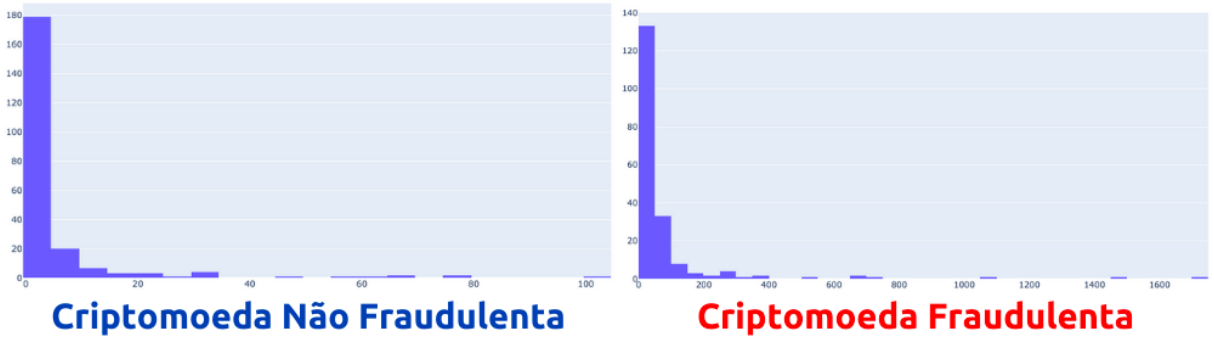


FIGURA 4.7 – Histograma de quantidade de dias de diferença entre as transações de um exemplo de criptomoeda fraudulenta e não fraudulentas

das fraudulentas acabarem não tendo tanto sucesso no mercado, pois acabam, no momento em que a fraude é descoberta, tendo uma queda brusca no fluxo de transações. Portanto, a quantidade total de transações de cada criptomoeda é uma medida importante, fora da série temporal em si.

Na Figura 4.8, observa-se o número de transações, onde cada ponto representa uma criptomoeda. Foram plotados os pontos em si, e também o boxplot de resumo estatístico. Neste gráfico, é notável que as criptomoedas não fraudulentas possuem muito mais transações do que as fraudulentas. Inclusive, entre os boxplots, não há *overlap*, o que indica robustez estatística nesta distinção.

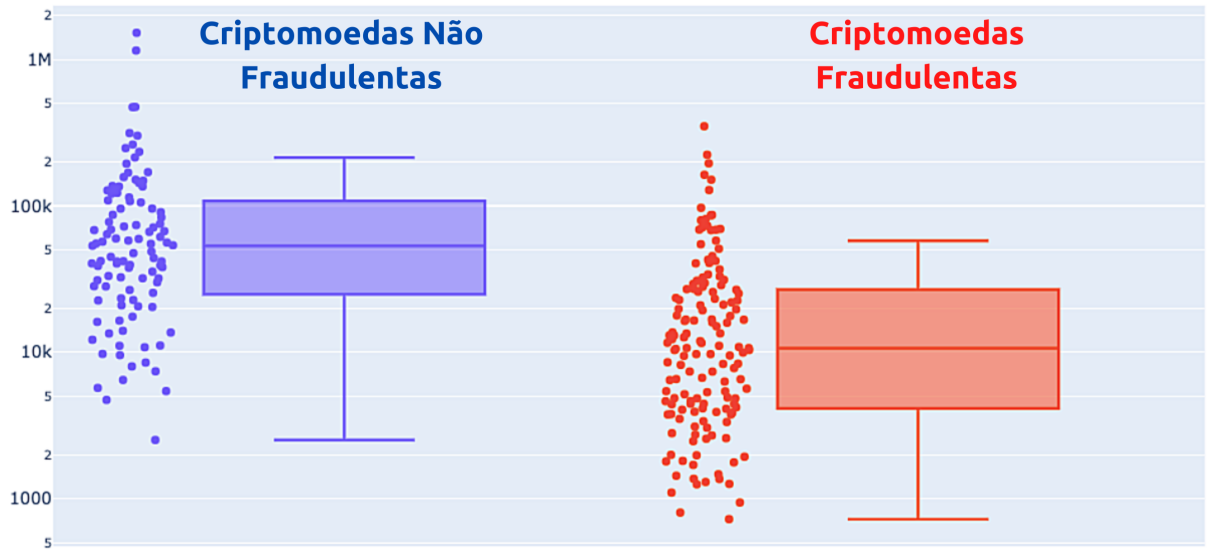


FIGURA 4.8 – Quantidade total de transações em seis meses de criptomoedas fraudulentas e não fraudulentas

4.3.5 Médias de GAS e GASLIMIT por transação

Uma das premissas elencadas anteriormente é que o papel da taxas das transações pode ser importante para detectar fraudes. Para tanto, calculou-se a média de GAS e GASLIMIT por transação utilizada para a mineração dos blocos para cada conjunto de dados dos 238 *datasets*.

Entretanto, as Figuras 4.9 e 4.10 não apresentam diferenças significativas. No caso do GASLIMIT, as criptomoedas não fraudulentas possuem uma média maior do que as fraudulentas. No caso do GAS, é o contrário.

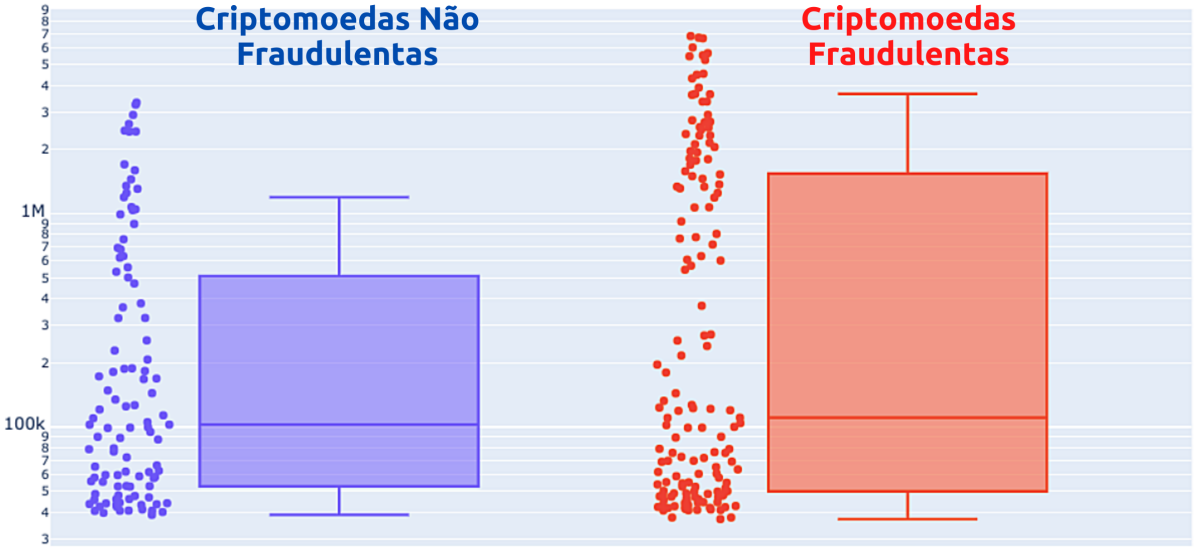


FIGURA 4.9 – Média de GAS por transação em seis meses de criptomoedas fraudulentas e não fraudulentas

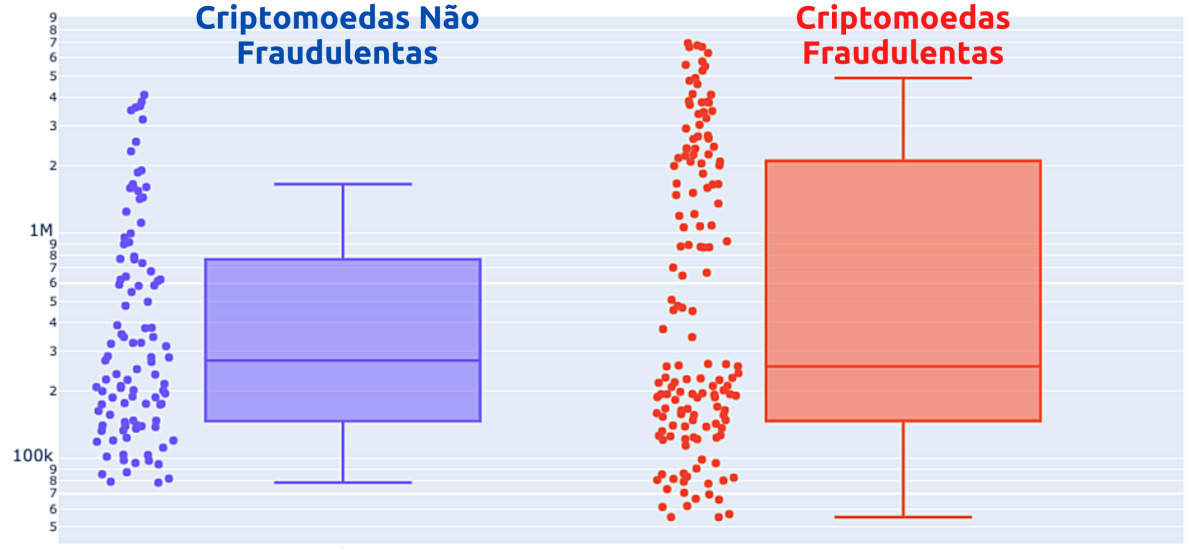


FIGURA 4.10 – Média de GASLIMIT por transação em seis meses de criptomoedas fraudulentas e não fraudulentas

4.3.6 Média de Vendedores Únicos de Transações

Vendedores únicos (*From Address*) é maneira de encontrar endereços distintos que tiveram interesse nas criptomoedas, em transações entre "From Address" e o "Smartcontract", por exemplo. É premissa que as criptomoedas fraudulentas tenham menos endereços distintos e que sejam controladas por *bots* que reaproveitam o mesmo endereço. Neste sentido, calculou-se a média da quantidade de transações de vendedores únicos por total de transações para cada *dataset*.

A Figura 4.11 permite observar que as criptomoedas são fraudulentas apresentam um comportamento de venda de *tokens* em que os vendedores são mais diversos, ao passo que nas criptomoedas fraudulentas, os vendedores tendem a se repetir com mais frequência. Isto quer dizer que, em um cenário onde não há fraude, há uma maior ocorrência de vendedores espontâneos.

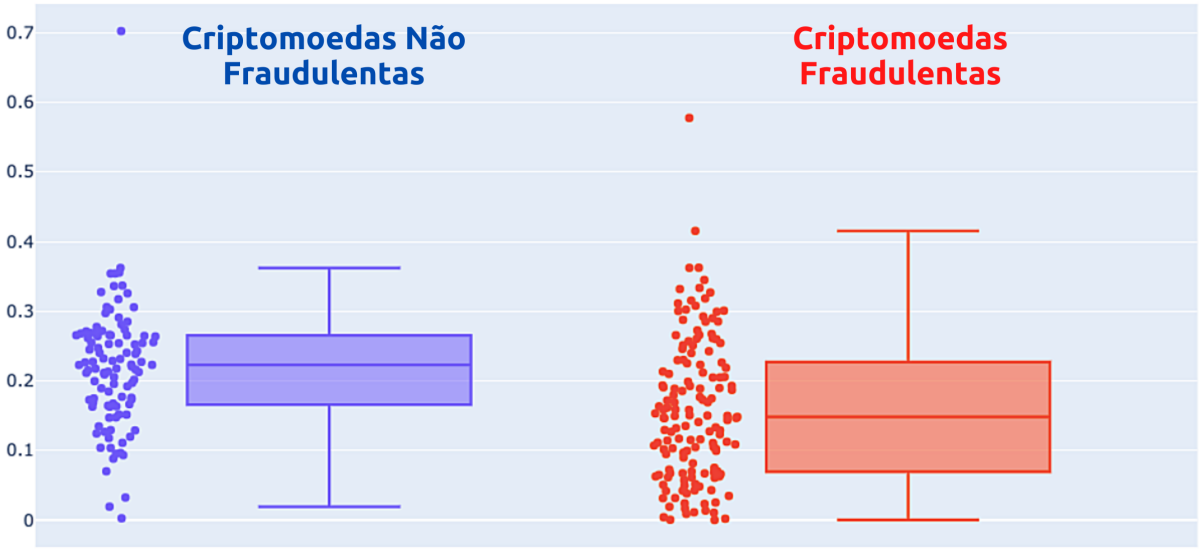


FIGURA 4.11 – Razão total de vendedores únicos por número de transações em seis meses de criptomoedas fraudulentas e não fraudulentas

4.3.7 Média de Compradores Únicos de Transações

Por outro lado, do ponto de vista de Compradores únicos (To Address), uma medida parecida com a anterior. Observando-se os vendedores, há um comportamento oposto ao verificado no item anterior. Desta forma, a média de transações de compradores únicos das criptomoedas por número de transações pode ser visualizada na Figura 4.12, onde há mais compradores únicos em criptomoedas fraudulentas do que em não fraudulentas. Uma explicação para isto é a presença de esquemas *pump and dump* em criptomoedas fraudulentas, onde os compradores automáticos são mal intencionados ao comprar títulos, com o objetivo de gerar um fluxo de transações, dando uma falsa impressão de que a

criptomoeda está sendo bem aceita no mercado.

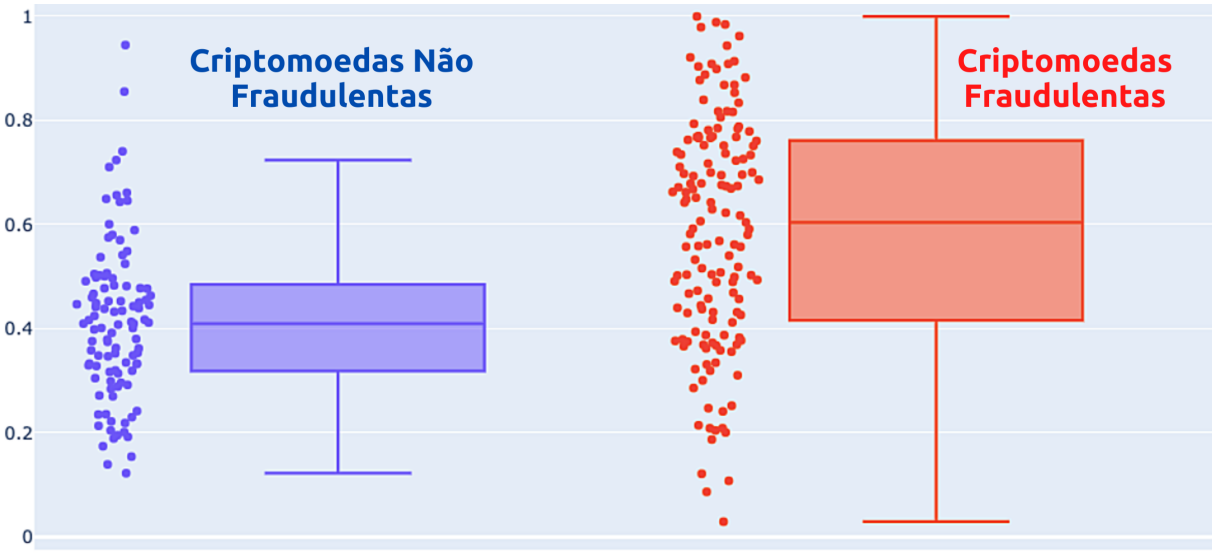


FIGURA 4.12 – Razão total de compradores únicos por número de transações em seis meses de criptomoedas fraudulentas e não fraudulentas

4.3.8 Média de transações de usuários novos na rede Ethereum

Por fim, foi calculada a média de transações de novos usuários das criptomoedas por total de transações, de acordo com a Figura 4.13. Nota-se que criptomoedas fraudulentas possuem a tendência de apresentar uma quantidade relativa maior de novos usuários na rede Ethereum. De fato, são pessoas ou *bots* que criam contas somente para executar atividades fraudulentas.

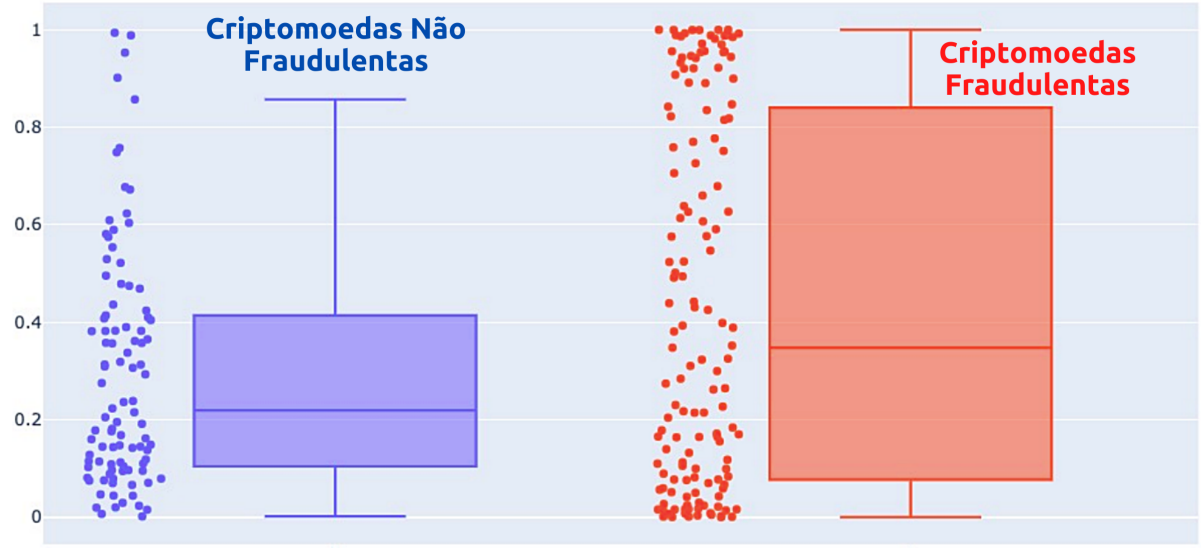


FIGURA 4.13 – Média de novos usuários em seis meses de criptomoedas fraudulentas e não fraudulentas

No escopo desta pesquisa, um novo usuário de uma criptomoeda é aquele que fez alguma transação cujo NONCE seja igual a “0” ou “1”. O NONCE igual a “1” também foi considerado porque pode ser que a primeira transação de um determinado usuário tenha sido a criação do *smart constract* e, por sua vez, este tipo de transação não está descrita nas tabelas CSV, uma vez que elas contém somente as transferências de criptomoedas de uma conta a outra.

4.4 Montagem das Séries Temporais

Esta Seção aborda o processo de montagem das cinco séries temporais, que serão usadas como entradas no modelo de classificação, baseado em RNA. O processo está de acordo com a Figura 4.14, no qual as entradas são as hipóteses da pesquisa e o comportamento dos seus dados e a saída é composta por pelas séries temporais.

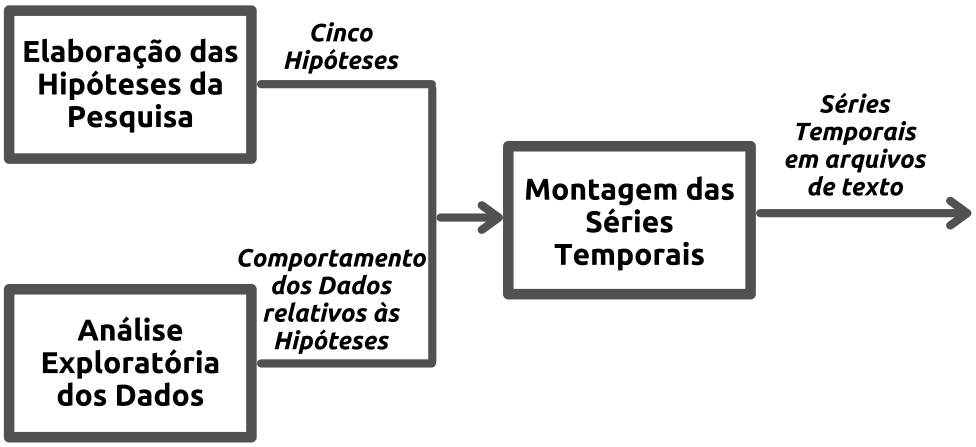


FIGURA 4.14 – Processo de Montagem das Séries Temporais.

4.4.1 Verificação de Condições para Classificação

Conforme abordado na Fundamentação Teórica, existem certas condições que permitem que as séries temporais possam ser usadas nos modelos de classificação baseados em RNA. Para este objetivo, realizou-se a consulta aos questionamentos de validação conforme Tabela 4.3.

A Figura 4.15, conforme já descrito na tabela de verificação, apresenta gráficos de auto-correlação de amostras de algumas criptomoedas classificadas como fraudulentas e não-fraudulentas. As auto-correlações são calculadas nas janelas de tempo de 1 dia, nos primeiros 50 dias.

TABELA 4.3 – Verificação das Séries usando Metodologia de (BROWNLEE, 2018).

Quais são as entradas e as saídas desejadas?	As entradas são as tabelas CSV obtidas pelo Google Big Query. As saídas são valores correspondentes a cada intuição lançada.
As variáveis são endógenas ou exógenas?	De acordo com o gráfico de autocorrelação, conforme a Figura 4.15, há séries que possuem comportamento de exógenas e outras de endógenas, pois a correlação entre as observações passadas varia de acordo com o gráfico de cada criptomoeda. Sendo assim, a priori, será considerado neste estudo que não há correlação entre as observações passadas. Sendo assim, as séries serão tratadas como exógenas.
O modelo é de classificação ou de regressão?	O modelo é de Classificação (criptomoeda fraudulenta ou não fraudulenta).
As séries são estruturadas ou não estruturadas?	As séries não são estruturadas. Entretanto, observa-se um número alto de quantidade de transações logo após a data de entrada da criptomoeda no mercado. Este número cai gradativamente ao longo do tempo.
O problema é de análise univariada ou multivariada?	O problema é de análise univariada e multivariada (as séries serão combinadas em uma mesma entrada).
A previsão se refere a um passo ou múltiplos passos ao longo do tempo?	O problema é de Classificação. Somente será observado se houve acerto ou não. O *Recall* será utilizado como o medidor de qualidade do modelo de Classificação.
O modelo de previsão é estático ou é dinamicamente atualizado?	O modelo de Classificação é estático.
As observações são contínuas ou descontínuas?	As observações são contínuas. De acordo com o histograma do espaço de tempo entre as transações, foi definido que o espaço de tempo padrão (utilizado no eixo X), ao longo deste trabalho, será de 01 dia. O início do eixo X será equivalente à data de entrada em que a criptomoeda entrou no mercado. E, por sua vez, o último número do eixo X é 60 dias após a data em que ela entrou no mercado.

4.4.2 Criação das Séries Temporais

Esta Seção apresenta como foram criadas as séries temporais para treinamento e testes nas redes neurais artificiais. Cada série temporal é baseada nas intuições explicadas no início deste Capítulo, e recebe o acrônimo do tipo de dados manipulado, conforme a Tabela 4.4.

No total, foram quatro séries criadas a partir dos Trabalhos Relacionados. Um dos objetivos deste trabalho é compará-las, a fim de verificar qual apresenta o melhor desempenho por ocasião da detecção de fraudes. Adicionalmente, foi desenvolvida uma séries TRANSACTIONS, com os dados não influenciados por nenhum Trabalho Relacionado.

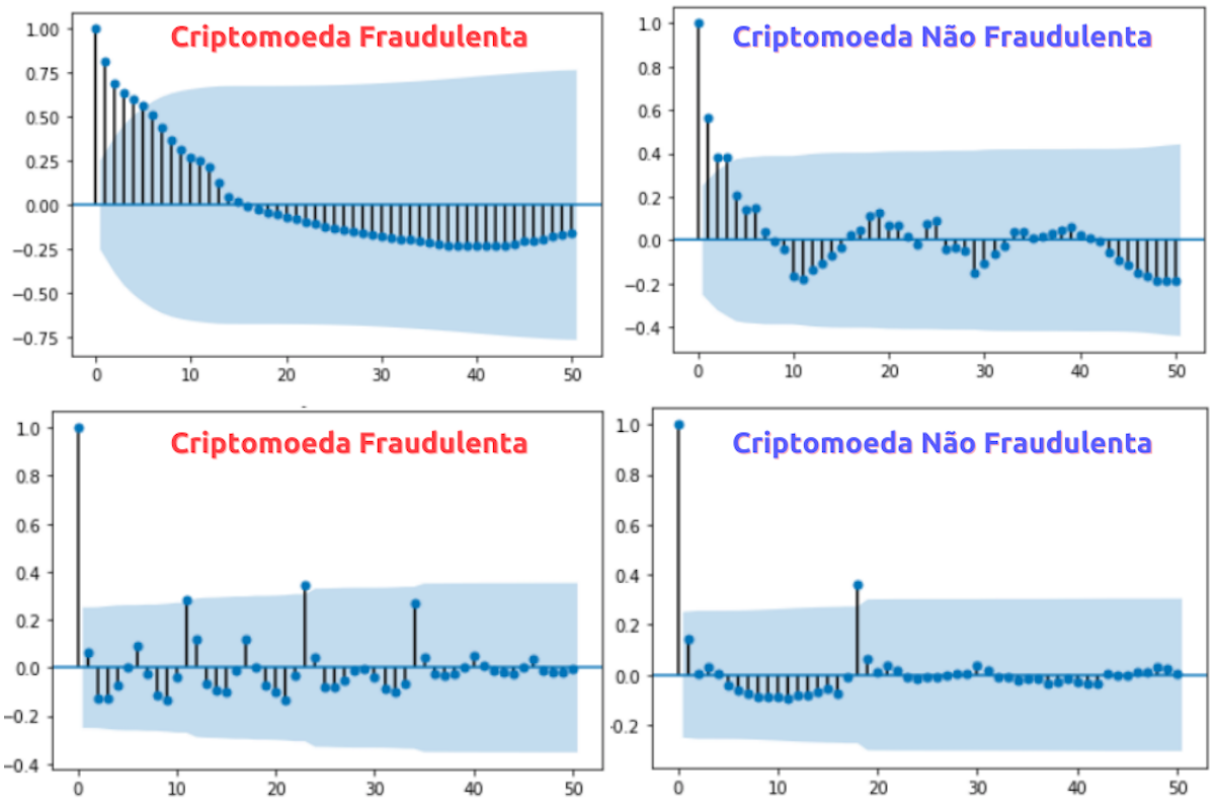


FIGURA 4.15 – Gráfico de autocorrelação entre dois exemplos de criptomoedas fraudulentas e dois de não fraudulentas

O propósito é comparar as quatro séries criadas com esta série, para verificar se os conhecimentos dos Trabalhos Relacionados são, de fato, relevantes para a montagem das séries temporais.

Cada uma das séries possui duas figuras que ilustram todas as séries relativas às criptomoedas fraudulentas (em vermelho) e todas as séries relativas às criptomoedas não fraudulentas (em azul). Adicionalmente, a primeira figura é atinente aos dados de valores brutos, ao passo que a segunda figura contem os dados suavizados, com uma média móvel relativa ao valor atual e duas observações passadas, todos com o mesmo peso.

TABELA 4.4 – Correlação entre as hipóteses e as séries temporais.

Hipótese	**Série Temporal**
NEWHOLDER	NEWHOLDER
NEWUSER	NEWUSER
BIGHOLDER	BIGBUYER
GAS/GASLIMIT	GAS/GASLIMIT
MARKETDATE	*Janelas de Tempo de 20, 40 e 60 dias*
-	TRANSACTIONS

Além disso, todas as séries estão representadas por linhas mais fracas, com os valores normalizados entre "0" e "1". Em cada figura, há duas linhas com cores mais fortes. A azul representa a média aritmética dos valores das criptomoedas não fraudulentas e a vermelha, a média aritmética dos valores das criptomoedas fraudulentas.

Cada Série Temporal possui as janelas de 20, 40 e 60 dias, a fim de seguir a hipótese MARKETDATE, a qual estabelece que a janela de tempo entre a data de entrada da criptomoeda no mercado e a data da sua análise é um fator relevante para a detecção de fraudes. Foram escolhidas estas três janelas de tempo, com o espaço de tempo igual entre elas (20 dias). Cada figura das séries a seguir são relativas às séries com janela de tempo de 60 dias.

4.4.2.1 Série Temporal NEWHOLDER

Tendo como objetivo verificar a hipótese NEWHOLDER, vinculada ao número de novos titulares ao longo do tempo, foi feita a modelagem desta Série Temporal NEWHOLDER, de acordo com a Equação 4.1 como o valor de referência no eixo-y.

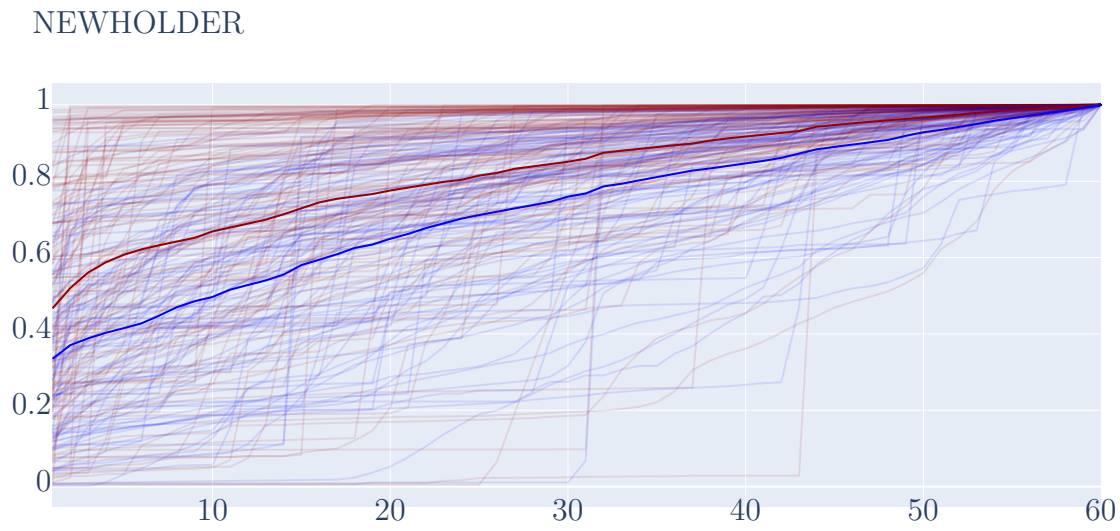
Para calcular o número total de participantes, considera-se o somatório de todos os usuários das colunas "FROM_ADDRESS" e "TO_ADDRESS" até 60 dias. Logo após, basta calcular o número de usuários que não se repetem. Seguindo a mesma lógica, para calcular o número de participantes no dia 32, por exemplo, basta fazer o mesmo procedimento, sendo que, ao invés de ir até 60, deve-se ir até 32.

$$y = \frac{\textit{número de participantes atual}}{\textit{número total de participantes}} \tag{4.1}$$

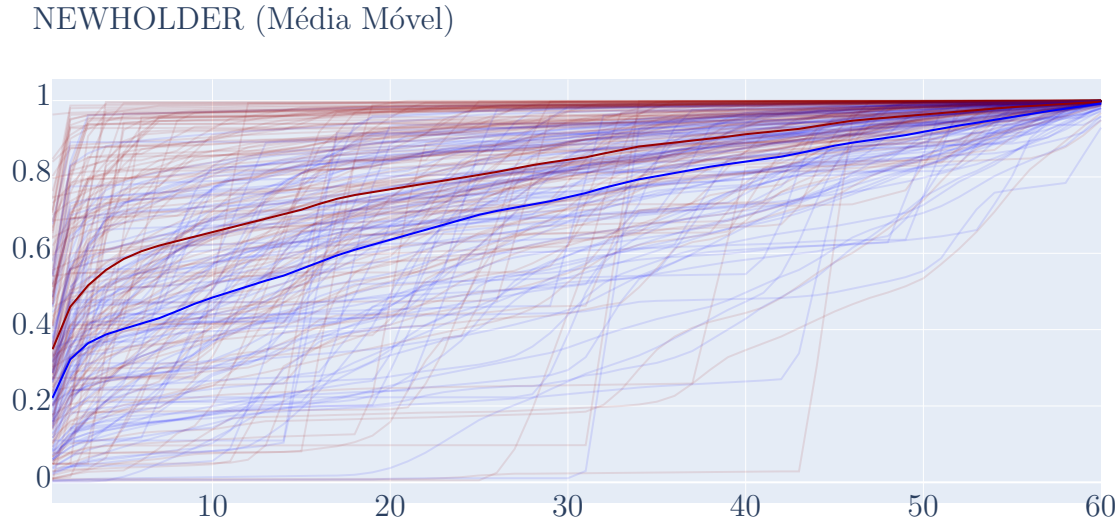
Modelando-se a Equação 4.1 dessa forma, a Série Temporal tem seus valores crescentes e cumulativos entre "0" e "1", sendo que o último número sempre terá seu valor igual a "1".

Nas Figuras 4.16a e 4.16b, mostramos lado a lado, todos os *datasets* utilizando essa métrica da S.T. NEWHOLDER de maneira pura, e ao lado, foram feitas as mesmas séries temporais, com apoio de uma média móvel para suavizar flutuações.

Além disso, há uma diferenciação entre as linhas mais fortes que expressam as médias aritméticas dos valores do eixo-y. Portanto, a expectativa é que o modelo tenha adequado desempenho ao classificar estas séries.



(a) Dados Brutos



(b) Dados Suavizados

FIGURA 4.16 – Séries Temporais representando, ao longo de 60 dias, o percentual de transações de usuários recém criados com suas médias

4.4.2.2 Série Temporal NEWUSER

A segunda série está relacionada com a quantidade de transações de usuários "recém-criados". Com o objetivo de incorporar esta formulação de usuário, desenvolveu-se a seguinte Equação 4.2 para modelar o eixo-y dessa série temporal:

$$y = \frac{número\ diário\ de\ transações\ de\ novos\ participantes}{número\ diário\ de\ transações} \tag{4.2}$$

Na Equação, há o número diário de transações de novos participantes. O NONCE é o mecanismo de proteção, como um contador, para evitar ataques de *replay* de mensagens. E, desta forma, é feita a filtragem na tabela de transações que contenha somente novos usuários. Por exemplo, supondo que no dia 25, no *dataset*, apareçam 24 transações de usuários novos e 48 transações diárias, o valor do eixo-Y será 0,5 apresentando então uma proporção de usuários novos em relação aos demais. A saída da função, portanto, já é normalizada.

As Figuras 4.17a e 4.17b apresentam esses resultados, novamente utilizando no lado A os dados brutos e no lado B os dados suavizados por uma média móvel. E as linhas fortes representam as médias dos conjuntos classificados como fraude e não-fraude. É possível observar que o poder de diferenciação parece ser menor nesta série, com os valores de fraude e não fraude estarem muito próximos de 0.2. Mas as criptomoedas fraudulentas tem um valor maior, conforme a hipótese, provavelmente pelo uso de *bots* para transmitir um aparente sucesso a esta criptomoeda.

4.4.2.3 Série Temporal BIGBUYER

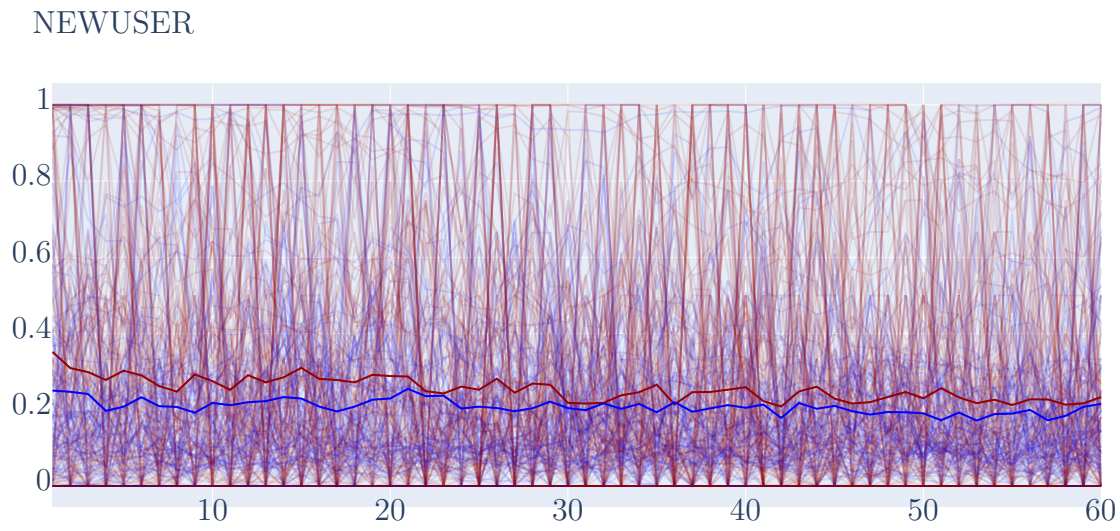
A próxima abordagem é relacionada à quantidade relativa do maior comprador de títulos de uma criptomoeda. Esta série temporal foi modelada conforme apresenta a Equação 4.3. A hipótese é que os títulos em criptomoedas fraudulentas fiquem na mão do fraudador que, por sua vez, quer atingir o máximo de preço para então vender todos os seus ativos.

No entanto, ao tentar encontrar o maior detentor de títulos das criptomoedas, ao longo do tempo, verificou-se que há incompatibilidade de dados, caso se use as tabelas de transações das criptomoedas. Uma possível explicação para isto é que há usuários que adquirem ativos, sem comprá-los. Sendo assim, esta aquisição não é registrada na tabela de transações da criptomoeda. Caso se queira desenvolver uma série que indique a porcentagem do usuário maior detentor de títulos ao longo do tempo, é preciso analisar os códigos dos *smart contracts*, um a um, o que tornaria esta pesquisa inviável, em detrimento do tempo.

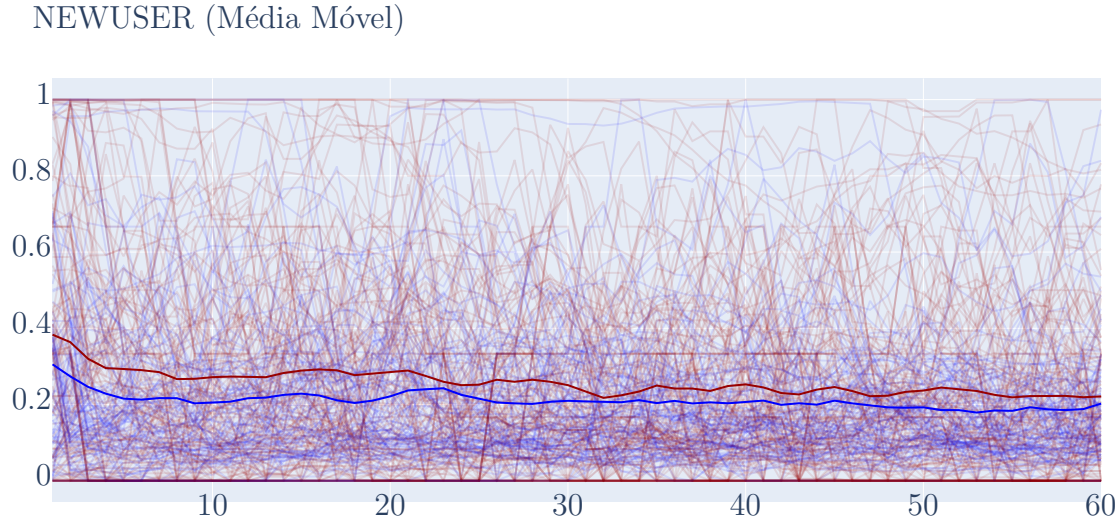
Dessa forma, a hipótese BIGHOLDER deu origem à série BIGBUYER, conforme a Equação 4.3:

$$y = \frac{Saldo\ do\ Maior\ Comprador\ de\ T\acute{i}tulos}{Total\ de\ T\acute{i}tulos\ Circulante} \tag{4.3}$$

Para o cálculo do maior comprador de títulos e número de títulos total, foi desenvolvida uma função "balanço" que rastreia pelo endereço ao longo do tempo, quantos tokens estão de posse daquele endereço. A implementação completa está contida no repositório



(a) Dados Brutos



(b) Dados Suavizados

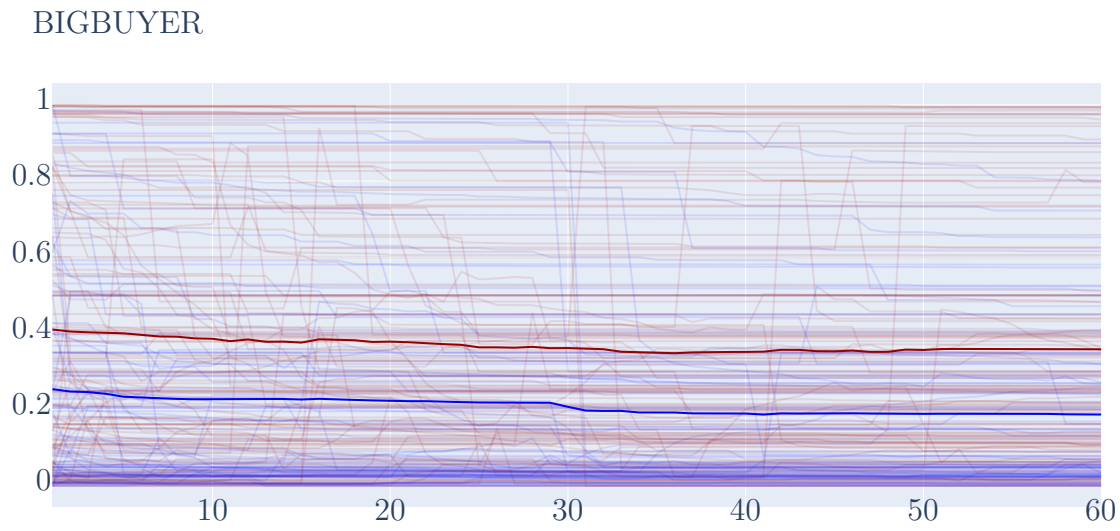
FIGURA 4.17 – Séries Temporais representando, ao longo de 60 dias, o percentual de transações de usuários recém criados com suas médias

Github[14].

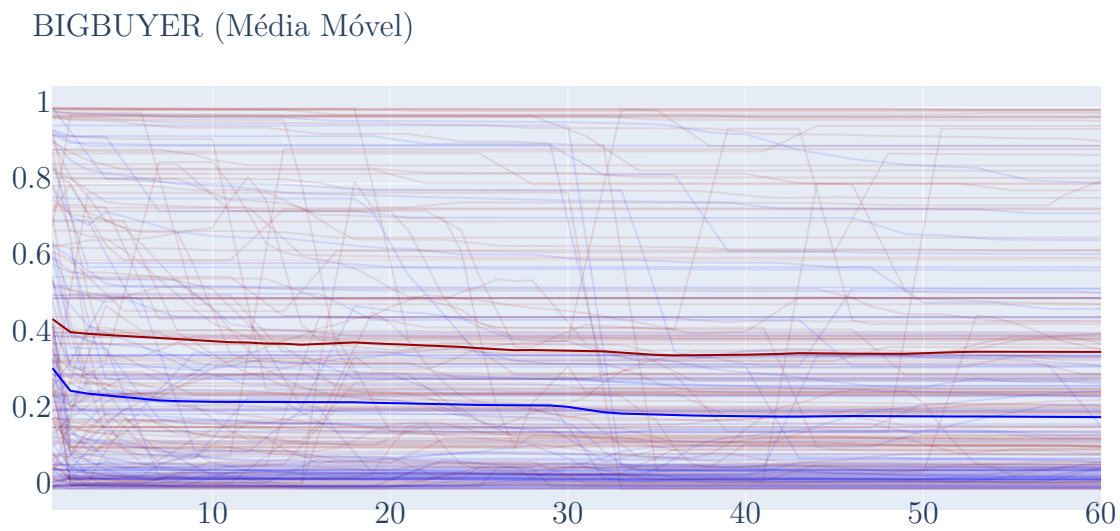
Nas Figuras 4.18a e 4.18b, pode-se observar o comportamento desta série do ponto de vista de todos os *datasets*. As médias gerais dos dois grupos de classificação apresentam uma diferenciação apropriada, sendo que as criptomoedas fraudulentas têm uma fração maior de usuários com ampla posse (linha vermelha versus azul).

Um aspecto importante sobre a criação desta série é que o maior comprador de títulos

[14]https://github.com/luizzmata/ICOFraudDetection/



(a) Dados Brutos



(b) Dados Suavizados

FIGURA 4.18 – Séries Temporais representando o percentual de títulos que cada maior comprador de títulos possui com suas médias

não pode ser um contrato, seguindo a inspiração da hipótese BIGHOLDER. Desta forma, foi necessária uma constante verificação, na rede Ethereum, se determinado usuário é um contrato ou não. Foi preciso utilizar-se da API da plataforma Etherscan para este objetivo. Todavia, a API limita-se a somente 5 requisições por segundo, tornando o tempo de confecção desta série maior do que o das outras presentes neste estudo.

4.4.2.4 Série Temporal GAS/GASLIMIT

Com base na hipótese de que o ambiente de taxas de transações (mediadas pelo GAS) pode ser manipulado para obter uma fraude com menor custo possível, foi estudada a série temporal com este parâmetro. Como ambos os parâmetros de GAS e GASLIMIT são intrinsecamente relacionados pelo projeto de aplicações em Ethereum, desta forma, o valor de referência pode ser dado como a razão entre elas, sendo o GASLIMIT o valor teto superior e o valor GAS o dado instantâneo. Portanto, a Equação 4.4 define para cada *dataset* este valor do eixo-y:

$$y = \frac{Total\ de\ GAS\ no\ dia}{Total\ de\ GASLIMIT\ no\ dia} \tag{4.4}$$

Naturalmente, o resultado da Equação é normalizado entre o valor GAS atual e o teto. Na Figura 4.19a, há variação desse parâmetro, tornando-se possível verificar visualmente, e também pelas linhas médias (linhas fortes), diferenciação entre elas.

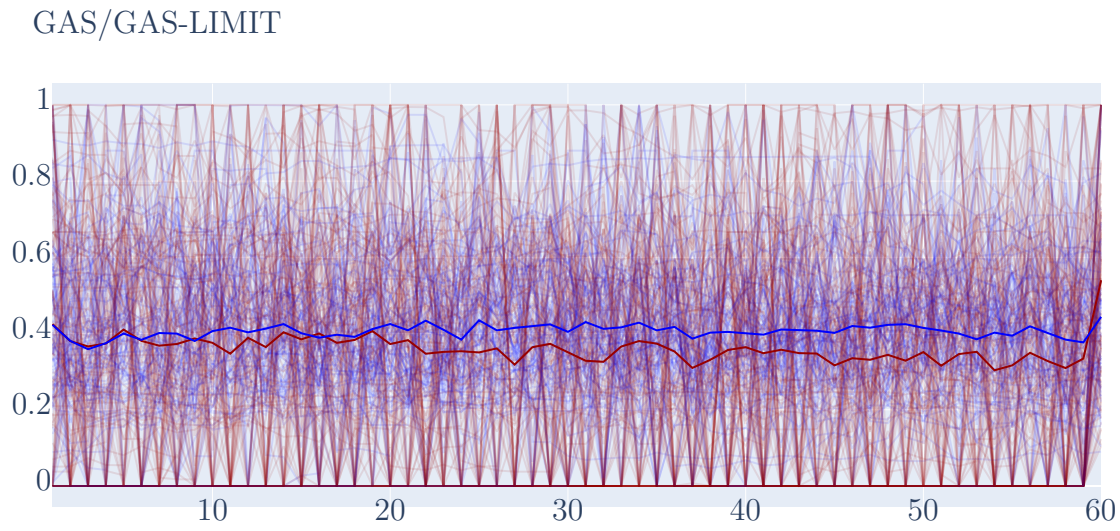
4.4.2.5 Série Temporal: TRANSACTIONS

Na área de previsão econômica, que tenta prever valores futuros de ações e assim obter lucros em boas compras e vendas no mercado financeiro, é comum o uso do valor mais simples, o preço da ação, para estimar o futuro. Do ponto de vista de atividades de ICO, o preço da criptomoeda é sempre o mesmo no período de criação. Portanto, o dado mais simples é o volume das transações. Deste modo, o total de transações por período tem o objetivo de ser usado de maneira comparativa com as séries específicas vistas anteriormente. A modelagem desta série é dada pela Equação 4.5:

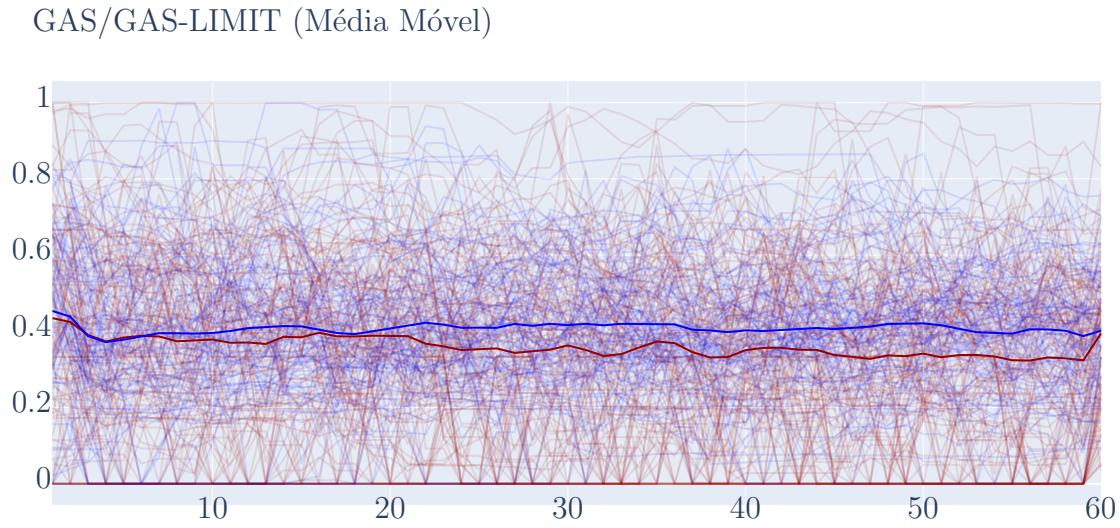
$$y = \frac{Quantidade\ de\ transações\ feitas}{Quantidade\ Total\ de\ transações} \tag{4.5}$$

Como pode ser observado na equação desta série, o valor é o cumulativo de transações em razão do tempo e o último valor do eixo-y é igual a "1". Por exemplo, se no dia 35, houver 4000 transações feitas até este dia e se o total de transações for 10.000, o valor de Y neste dia será de 0,4.

Como antes fora abordado, o desenvolvimento desta série não está vinculado a nenhum Trabalho Relacionado específico. Antes, é tão somente para fins comparativos, para verificar se os conhecimentos dos Trabalhos Relacionados, aplicados nas séries anteriores, são relevantes para a detecção de fraudes.



(a) Dados Brutos



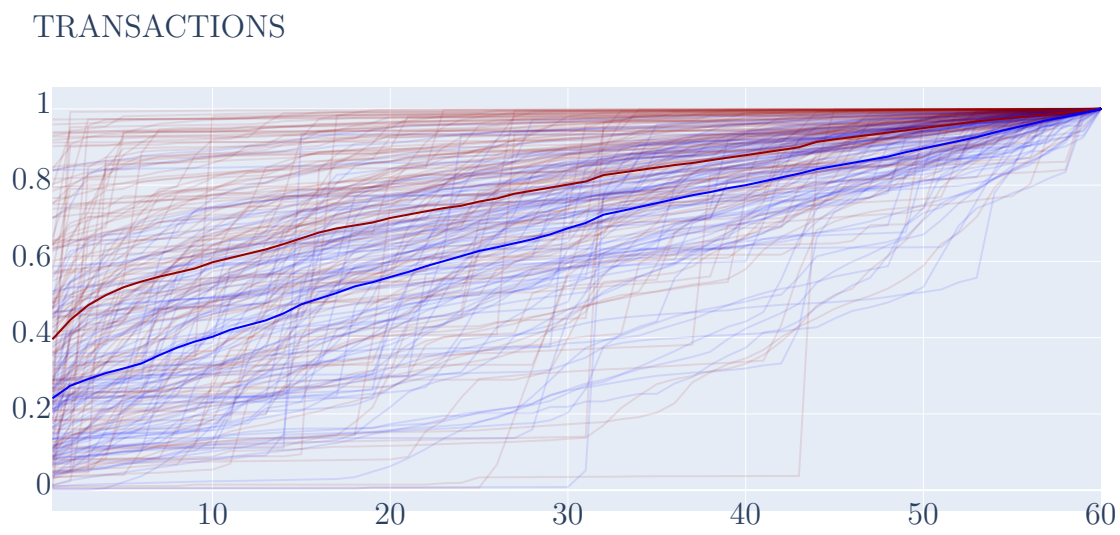
(b) Dados Suavizados

FIGURA 4.19 – Séries Temporais, ao longo de 60 dias, representando a razão GAS / GASLIMIT com suas médias

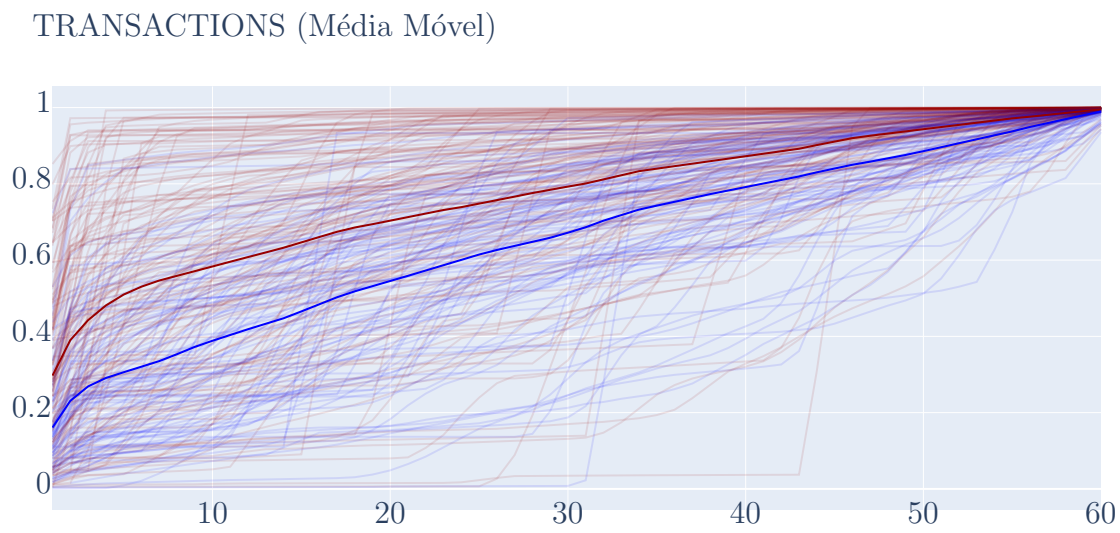
4.5 Montagem dos Modelos de Classificação em RNA

Nesta Seção, são abordadas: as descrições dos modelos de RNA utilizados para os experimentos e a metodologia empírica usada para a escolha de hiperparâmetros e obtenção dos resultados. Este é o último passo do Método de Detecção de Fraudes, descrito na Figura 4.21.

De acordo com o escopo deste estudo, o modelo de classificação é de RNA composta por um ou mais tipos de arquiteturas, destinadas a propiciar a classificação dos dados de



(a) Dados Brutos



(b) Dados Suavizados

FIGURA 4.20 – Séries Temporais representando o percentual do número de transações, ao longo de 60 dias, com suas respectivas médias aritméticas em destaque

entrada. No caso, os dados de entrada representam as séries temporais, em arquivos de texto, e a saída representa o desempenho do modelo em classificar as séries, medido em *Recall*.

4.5.1 Descrições dos Modelos de RNA

Foram projetados três modelos de RNA: MLP, CNN-MLP e LSTM-MLP. Estes modelos realizam o trabalho de classificar as séries temporais, com o objetivo de detectar frau-

FIGURA 4.21 – Processo de Montagem dos Modelos de RNA.

des. No total, são 15 séries temporais, uma combinação de 5 tipos de séries (NEWHOLDER, NEWUSER, BIGBUYER, GAS/GASLIMIT e TRANSACTIONS) com 3 janelas de tempo (20, 40 e 60 dias). No escopo deste estudo, experimento é a combinação das três variáveis (séries temporais, janelas de tempo e modelos de RNA), totalizando 45 experimentos.

4.5.1.1 Descrição dos *notebooks* Jupyter

Todos os experimentos foram realizados em um computador Intel 64 bits com 8 cores e 64GB de RAM, com disco SSD e 2 placas GPU NVidia 1080i (3000 cores). A suíte de desenvolvimento foi o Jupyter notebook, versão 6.0.3, instalado no Python, versão 3.7.3, com as bibliotecas tensorflow, keras, pandas e numpy.

Ao longo do trabalho, foram desenvolvidos nove *notebooks* Jupyter, cada um referente aos dados relativos a cada janela de tempo (*notebook* com os dados de 20 dias, outro, com os dados de 40 dias e o último, com os dados de 60 dias), combinados com cada modelo de RNA abordado neste estudo (MLP, CNN-MLP e LSTM-MLP). Para o treinamento das séries, está pré-configurado com o nome da série a ser treinada, conforme pode ser visualizado na Figura 4.22.

FIGURA 4.22 – Jupyter Notebook relativo à janela de 60 dias.

Os três modelos, bem como seus hiperparâmetros, encontram-se descritos na Ta-

TABELA 4.5 – Descrição dos modelos de RNA

Arquitetura	**Camada**	**Modelo**	**Nº**	**Função de Ativação**	**Demais Hiperparâmetros**
MLP	Entrada	Densa	60	ReLU	-
	Escondida	Densa	30	ReLU	-
	Escondida	Densa	20	ReLU	-
	Escondida	Densa	10	ReLU	-
	Saída	Densa	1	Sigmoid	-
CNN-MLP	Entrada	Convolucional	1	ReLU	Filtros = 8 Tamanho do kernel = 3
	Escondida	Convolucional	1	ReLU	Filtros = 8 Tamanho do kernel = 3
	Escondida	Dropout	1	-	0.5
	Escondida	MaxPooling	1	-	Tamanho do pool = 2
	Escondida	Achatamento	1	-	-
	Escondida	Densa	224	ReLU	-
	Saída	Densa	1	Sigmoid	-
LSTM	Entrada	LSTM	100	TanH Sigmoide	-
	Escondida	Densa	60	ReLU	-
	Saída	Densa	1	Sigmoid	-

bela4.5. Cada modelo de RNA contém uma camada de entrada, uma de saída e uma ou mais escondidas. A Função de ativação escolhida para as camadas, exceto a de saída, foi a *ReLU*. A função de ativação de saída final foi a *Sigmoid*.

Em relação especificamente à arquitetura CNN-MLP, foi necessário estabelecer hiperparâmetros para as camadas convolucionais, como quantidade de filtros, tamanho do *kernel* e tamanho do *pool*.

4.5.1.2 Treinamento e Predição

Os dados das séries temporais ficaram organizados em três arquivos de texto, a fim de otimizar o tempo para a criação das séries, por ocasião do treinamento e predição. Cada arquivo de texto contem, respectivamente, todos os valores das séries de 20, 40 e 60 dias.

Por ocasião do treinamento e predição, todas as criptomoedas foram divididas, aleatoriamente, entre dois conjuntos: conjunto de treinamento (70% dos dados) e conjunto de predição (30% dos dados). Desta forma, das 238 criptomoedas, 166 ficaram no grupo de treinamento e 72, no grupo de predição.

Cada uma das três arquiteturas em cada *notebook* Jupyter possui células de processamento específicas para executar o treinamento e predição, conforme a Figura 4.23

FIGURA 4.23 – Treinamento e Predição dos modelos de RNA

TABELA 4.6 – Hiperparâmetros dos Treinamento e Predição

Hiperparâmetro	**Valores**
Otimizador	Adam
Métrica de Desempenho	Recall
Épocas	100 para MLP e CNN-MLP 400 para LSTM-MLP
Tamanho do *Batch*	128, 64, 32 e 16

Todo o treinamento e validação possui os hiperparâmetros, apresentados na Tabela4.6. O otimizador, para todos, foi o Adam. A métrica utilizada para análise dos desempenhos foi o *Recall*. Ressalta-se que o *Recall* ou sensibilidade é definida como o número de previsões positivas verdadeiras (fraudulentas) em comparação com o número total de transações fraudulentas. Na detecção de fraudes, a medida mais importante é esta, pois maior *recall* significa menor perda financeira.

Em relação ao número de épocas, foi escolhida uma quantidade que pudesse unir otimização de tempo (a partir de um certo ponto, a RNA não melhora o índice de *Recall* ou entra em *overfitting*) à obtenção de máximo desempenho possível. Sendo assim, foi escolhido o número de 100 épocas para os modelos MLP e CNN-MLP e 400 épocas para o modelo LSTM-MLP. Foi observado que, neste último, devido a componente de memória longa, precisava-se de mais épocas para que o modelo pudesse alcançar um desempenho maior, sem entrar em *overfitting*.

Por último, foi escolhido o tamanho de *Batch* 128 como principal. Contudo, houve séries em que o *Recall* ficou estabilizado ao longo das épocas e, desta forma, foram utilizados, em ordem de preferência, os tamanhos 128, 64, 32 e 16, o que será visto com mais detalhes na próxima seção.

4.5.2 Método para Obtenção dos Resultados

Primeiramente, será preciso entender como identificar no gráfico de resultados os valores de *Recall* e perda (tanto no treinamento como na predição), quando ocorre *Overfitting*

e as implicações na modificação do hiperparâmetro "Tamanho do *Batch*". Logo após, será descrita a metodologia utilizada para a aquisição dos resultados.

4.5.2.1 Valores de *Recall* e perda

Os critérios de corte para a escolha do melhor modelo com a melhor parametrização e resultados finais foram obtidos a partir da visualização de gráficos cartesianos, contemplando número de épocas (eixo X) e valores de *Recall* e perda, tanto no treinamento, quanto na predição (eixo Y), conforme pode ser visto na Figura 4.24.

```
ico_training.train_network(loss='binary_crossentropy',
                      optimizer='adam',
                      metrics=[Recall()] ,
                      epochs=100,
                      verbose=0,
                      batch_size=128,
                      callback=earlystop)

ico_training.plot_training()
```

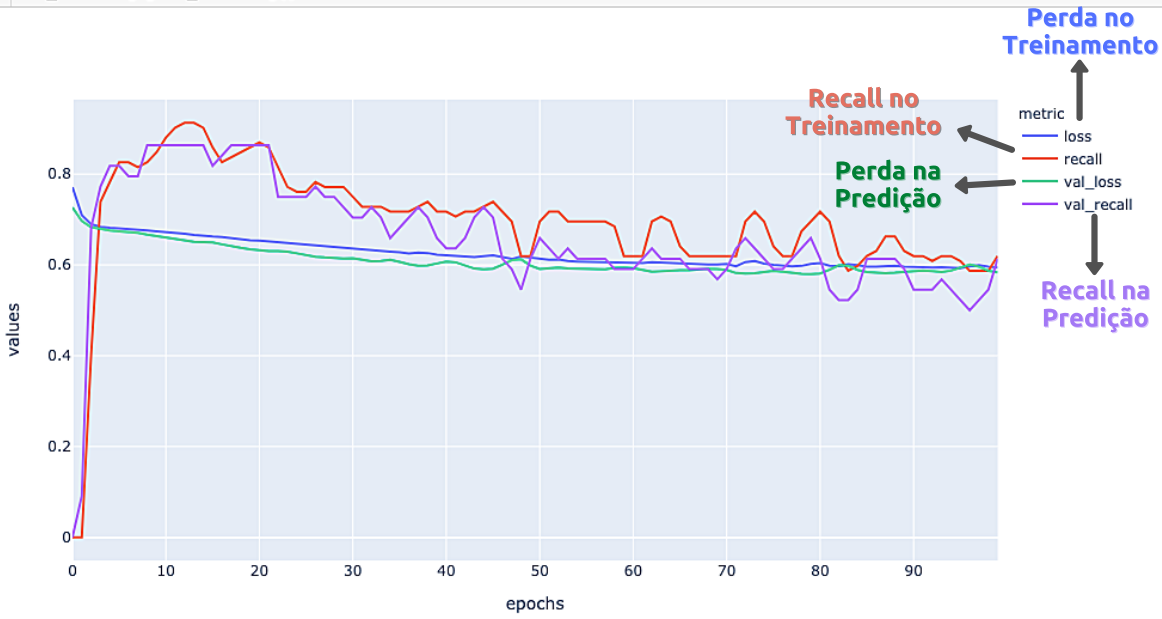


FIGURA 4.24 – Valores de *Recall* e perda nos treinamento e predição ao longo da quantidade de épocas.

Os valores de *Recall* da predição costumam estar um pouco abaixo dos valores de *Recall* do treinamento. Outrossim, os valores da perda de treinamento e predição seguem o mesmo comportamento.

4.5.2.2 Detectando *overfitting*

Outro aspecto importante é detectar o *overfitting*. Ele é notado visualmente, quando os valores de perda na predição tendem a se afastar continuamente dos valores de perda no treinamento. Adicionalmente, os valores de *Recall* de treinamento costumam ficar iguais

ou próximos a 100%, enquanto os valores de *Recall* da predição, próximos ou iguais a zero, de acordo com a Figura 4.25 e Figura 5.2.

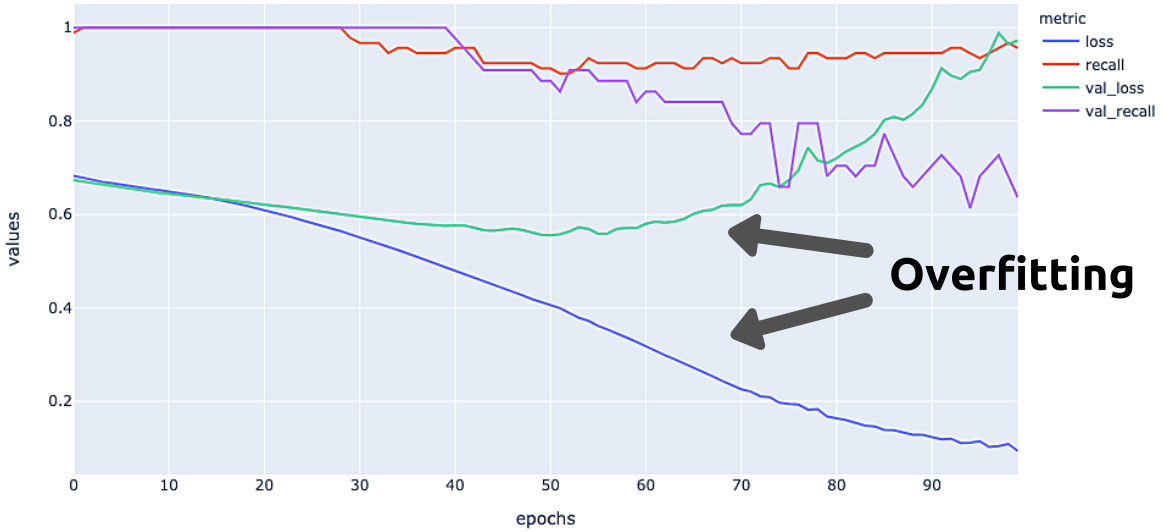


FIGURA 4.25 – Valores de perda no treinamento se distanciando dos valores de perda na predição, indicando *overfitting*.

4.5.2.3 Modificando o Tamanho do *Batch*

Foi observado que quando diminui-se o Tamanho do *Batch*, aumenta-se a frequência de oscilação dos valores de *Recall*, bem como sua amplitude, ao passo que, quando aumenta-se o Tamanho do *Batch*, diminui-se a frequência de oscilação dos valores de *Recall*, bem como sua amplitude. Ou seja, *Batch* mais amplo suaviza o treinamento. Tal efeito pode ser observado nas figuras 4.26a e 4.26b.

(a) **Tamanho do *Batch* = 16** (b) **Tamanho do *Batch* = 128**

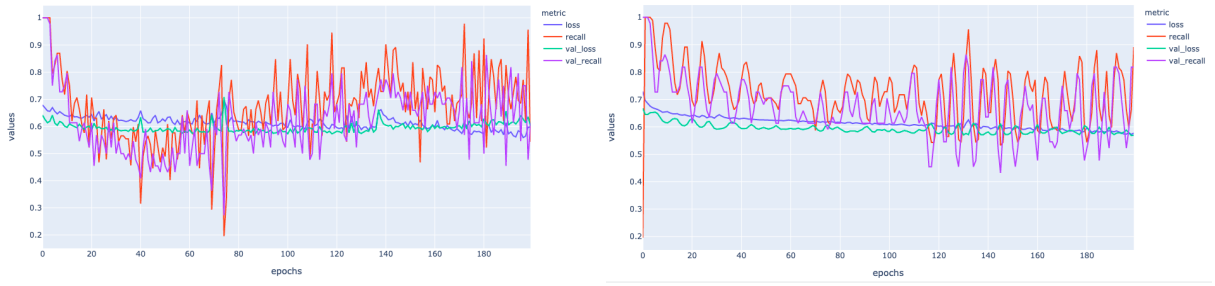


FIGURA 4.26 – Exemplo de diferença de comportamento dos valores de *Recall*, ao alterar o Tamanho do *Batch*.

Contudo, esta característica pode ser utilizada positivamente, a fim de se obter resultados melhores, por ocasião dos picos dos valores de *Recall*. As figuras 4.27a e 4.27b mostram uma comparação entre dois treinamentos e predições apenas variando o Tama-

nho do *Batch*. É possível observar que, em um Tamanho do *Batch* menor, há picos de desempenho maiores do que quando executando com um *Batch* maior.

(a) **Tamanho do *Batch* = 32** (b) **Tamanho do *Batch* = 128**

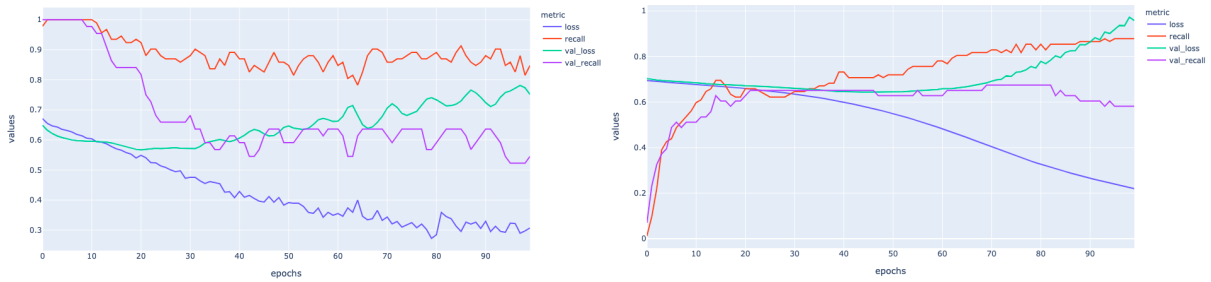


FIGURA 4.27 – Comparação entre dois Tamanhos do *Batch* diferentes.

4.5.2.4 Método para Aquisição dos Resultados

Uma vez introduzidos conceitos importantes, como a detecção de *Overfitting* e a influência do Tamanho do *Batch*, neste ponto, será abordado o método para a obtenção dos resultados, ou seja, um critério para a escolha dos melhores parâmetros. Esta importante escolha é composta por uma análise empírica do gráfico de desempenho do modelo de cada RNA, obedecendo aos seguintes passos:

1. Encontre o número de épocas em que é presenciado o *overfitting*;
2. Encontre o número de épocas igual ao Tamanho do conjunto de treinamento somado com validação (ou seja, 238) dividido pelo Tamanho do *Batch*. Tal medida tem a finalidade de certificar que todas as criptomoedas passaram pelo menos uma vez pelo treinamento ou predição;
3. Caso encontre o ponto da linha do *Recall* da Predição, entre o item 1 e item 2, que seja um pico de alta, este será o resultado;
4. Caso não encontre o ponto do item 3, diminua o Tamanho do *Batch*, sucessivamente, para 64, 32 e 16; e
5. Caso não encontre o ponto, ao passar pelo item 4, assumir que não possível treinar o modelo.

As Figuras 4.28a e 4.28b mostram um exemplo para o cálculo do desempenho da NEWUSER, modelo CNN-MLP, com 20 dias.

Primeiramente, foi realizado o procedimento de treinamento e predição com Tamanho de *Batch* 128. Foram considerados os resultados entre as épocas 2 (ponto em que todas

(a) **Tamanho do *Batch* = 128** (b) **Tamanho do *Batch* = 64**

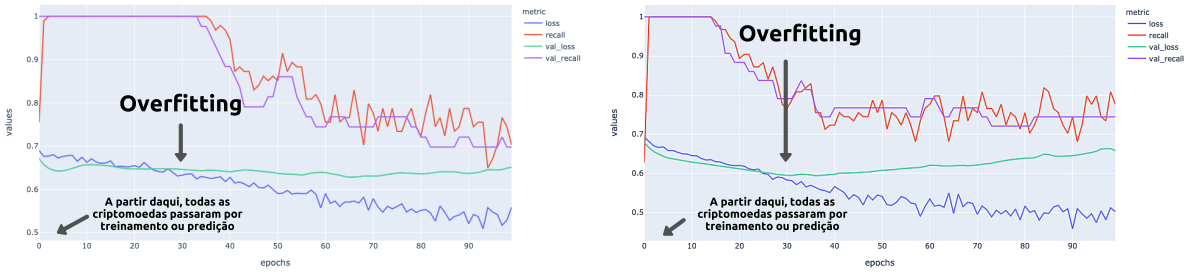


FIGURA 4.28 – Processo de obtenção do *Recall* para a NEWUSER, modelo CNN-MLP, com 20 dias.

as criptomoedas já treinadas ou preditas) e 23 (ponto em que iniciou-se o *Overfitting*. Contudo, os valores de *Recall* estavam ainda 100%. Sendo assim, foi reduzido o Tamanho do *Batch*.

Ao ser reduzido para 64 o Tamanho do *Batch*, foi observado que os valores de *Recall* saíram de 100% e se formaram picos, a partir da época 16. Adicionalmente, os dados considerados, da mesma forma que o experimento anterior foram entre as épocas 4 (ponto em que todas as criptomoedas já foram treinadas ou preditas) e 28 ponto em que se iniciou o *overfitting*. Foi considerado como valor do *Recall*, 83%, na época 25. Este procedimento foi realizado para cada um dos 45 experimentos abordados neste estudo.

4.6 Resumo do Método de Detecção de Fraudes

Foram vistos os cinco passos do método, conforme Figura 4.29. Primeiramente, foram elaboradas cinco hipóteses, baseadas em três Trabalhos Relacionados.

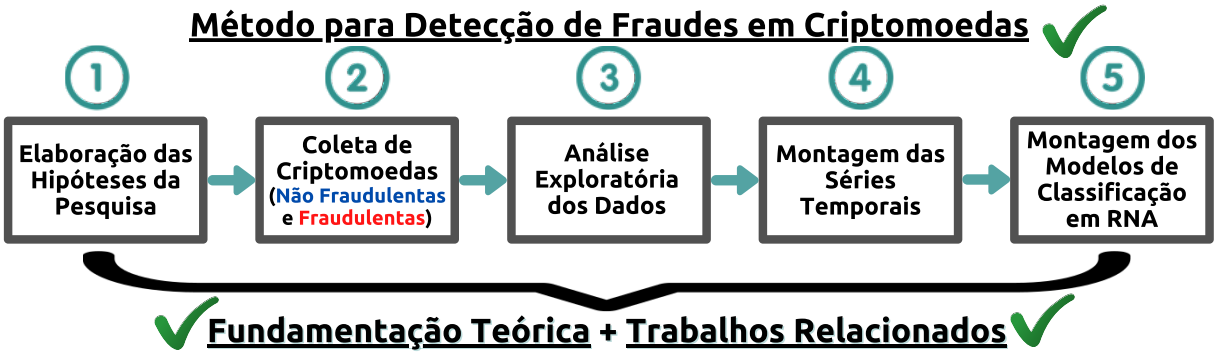


FIGURA 4.29 – Método para Detecção de Fraudes em Criptomoedas.

Em seguida, projetou-se o processo de coleta de criptomoedas fraudulentas e não fraudulentas, de acordo com critérios pré estabelecidos, sob direcionamento de três Trabalhos Relacionados. O resultado foi o conjunto de 238 tabelas CSV, 102 para criptomoedas não fraudulentas e 136, para fraudulentas.

Logo após, com base nesses *datasets*, realizou-se a Análise Exploratória dos Dados, a fim de investigar o comportamento das características previstas nas hipóteses, como fluxo de transações, novos usuário, GAS e GASLIMIT.

Então, sob as bases dos resultados desta análise e das hipóteses elaboradas, criou-se as quatro séries temporais de NEWHOLDER, NEWUSER, BIGBUYER e GAS/GASLIMIT. Adicionalmente, foi criada mais uma série (TRANSACTIONS), a fim de realizar uma comparação de desempenho de classificação entre as séries advindas dos Trabalhos Relacionados e uma série simples. Além disto, foram estabelecidas, para cada série, três janelas de tempo (20, 40 e 60 dias).

Por último, três modelos de RNA foram projetados. No total, foram realizados 45 experimentos, fruto de uma combinação entre cinco séries temporais, três janelas de tempo e três modelos de RNA. As entradas representaram as séries e a saída o desempenho, medido em *Recall*.

Finalmente, o Método de Detecção de Fraudes em Criptomoedas foi concluído. O próximo Capítulo apresenta os resultados obtidos.

5 Resultados dos Modelos de RNA

Neste Capítulo, encontra-se uma análise comparativa dos resultados, a fim de se verificar qual variável (série temporal, janela de tempo e modelo de RNA) é melhor para a detecção de fraudes.

Dessa forma, este Capítulo é dividido em quatro seções: Comparação dos resultados entre as séries temporais; Comparação dos resultados entre as janelas de tempo; Comparação dos resultados entre os modelos de RNA; e Síntese dos resultados.

Com o objetivo de ilustrar uma visão holística dos resultados, optou-se pela apresentação em *heatmap* (mapa de calor), no qual os valores máximo, médio e mínimo são, respectivamente: cor azul (68%), cor branca (79,5%) e cor vermelha (91%). Combinações formam degradê nas figuras apresentadas.

5.1 Comparação dos resultados entre as séries temporais

No total, como fora explicado no Capítulo 4, quatro séries foram criadas, a partir dos conhecimentos advindos dos Trabalhos Relacionados e uma série simples, indicando o crescimento da quantidade de transações ao longo do tempo.

Sendo assim, foi realizada uma comparação de resultados entre as séries criadas, a fim de identificar qual a melhor série, em termos de desempenho na classificação e uma comparação entre as séries criadas e a séries simples, TRANSACTIONS, a fim de verificar se a aplicação dos conhecimentos descritos nos Trabalhos Relacionados nos desenvolvimentos das séries contribuíram para a melhoria da detecção de fraudes.

5.1.1 Comparação entre as séries temporais desenvolvidas

Primeiramente, encontram-se descritos os resultados de cada série temporal desenvolvida (NEWHOLDER, NEWUSER, BIGBUYER e GAS/GASLIMIT), a partir dos conhecimentos adquiridos pela revisão da literatura existente sobre este tema. Os resultados estão apresentados na a Figura 5.1 e estão comentados na Tabela descrita na Figura 5.1.

Redes / Séries	NEWHOLDER			NEWUSER			BIGBUYER			GAS/GASLIMIT		
	20 dias	40 dias	60 dias	20 dias	40 dias	60 dias	20 dias	40 dias	60 dias	20 dias	40 dias	60 dias
MLP	83%	76%	86%	83%	86%	88%	86%	86%	88%	84%	81%	88%
CNN-MLP	86%	82%	81%	83%	88%	90%	86%	86%	79%	75%	81%	88%
LSTM-MLP	83%	91%	77%	91%	88%	75%	80%	83%	75%	-	-	-

FIGURA 5.1 – Análise comparativa de métrica *Recall* pelo tipo de série criada

TABELA 5.1 – Comparação de resultados das séries criadas

Séries	Observações
NEWHOLDER	Foi alcançado o desempenho de 80% ou mais em 7 de 9 testes. Destaca-se o modelo CNN-MLP, onde atingiu o resultado de 91% em 40 dias após a data de entrada da criptomoeda no mercado.
NEWUSER	Foi alcançado o desempenho de 80% ou mais em 8 de 9 testes. Destaca-se o modelo LSTM-MLP, onde atingiu o resultado de 91% em 20 dias após a data de entrada da criptomoeda no mercado.
BIGBUYER	Foi alcançado o desempenho de 80% ou mais em 7 de 9 testes. Destaca-se o modelo MLP, onde atingiu o resultado de 88% em 60 dias após a data de entrada da criptomoeda no mercado.
GAS/GASLIMIT	Foi alcançado o desempenho de 80% ou mais em 5 de 9 testes. Destaca-se o modelo MLP e CNN-MLP, onde atingiram o resultado de 88% em 60 dias após a data de entrada da criptomoeda no mercado. O modelo LSTM-MLP não conseguiu treinar essas séries, tendo em vista o constante *overfitting* detectado, conforme figura 5.2.

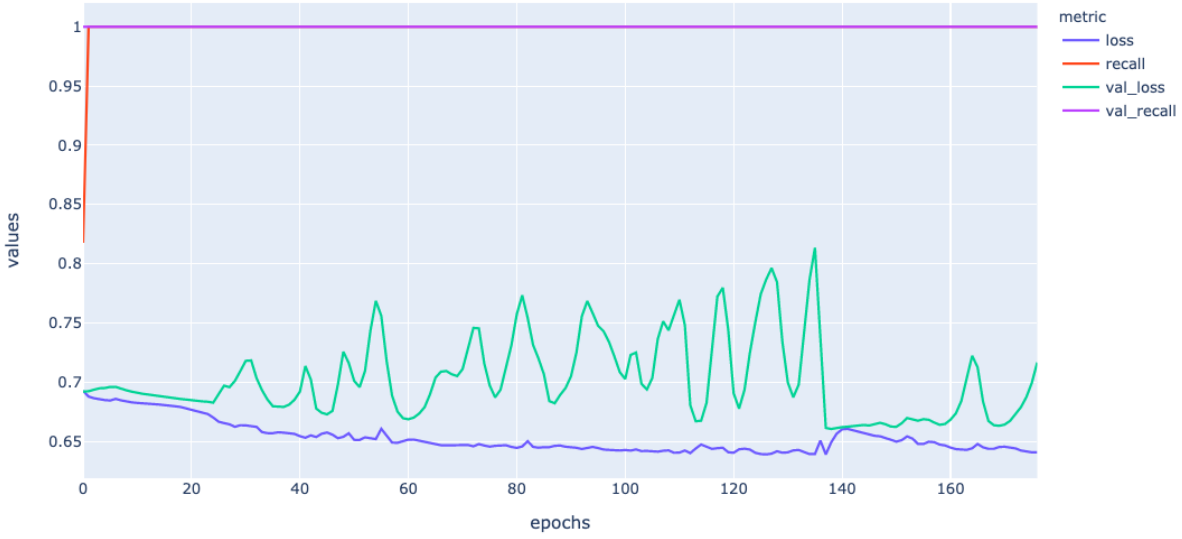


FIGURA 5.2 – *Overfitting detectado na série GAS por GASLIMIT ao passar pelo modelo LSTM-MLP*

TABELA 5.2 – Análise Comparativa entre a média de desempenho entre a série Número de Transações e a média aritmética do *Recall* das séries criadas

NEWHOLDER	NEWUSER	BIGBUYER	GAS/GASLIMIT	TRANSACTIONS
82,8%	85,8%	83,2%	82,8%	79,1%

5.1.2 Comparação entre as séries criadas e a série TRANSACTIONS

Foi realizada uma comparação entre a média aritmética dos resultados das séries desenvolvidas e os resultados da série TRANSACTIONS, com o objetivo de mostrar qual a relevância dos conhecimentos obtidos, por meio da revisão da literatura, na criação de séries temporais mais elaboradas. A Figura 5.3 descreve os resultados.

Redes / Séries	TRANSACTIONS			Média das séries criadas		
	20 dias	40 dias	60 dias	20 dias	40 dias	60 dias
MLP	81%	76%	72%	84,0%	82,3%	87,5%
CNN-MLP	88%	82%	68%	82,5%	84,3%	84,5%
LSTM-MLP	88%	85%	72%	84,7%	87,3%	75,7%

FIGURA 5.3 – Análise comparativa de métrica *Recall* do número de transações com a média das séries criadas

Como pode ser visto, os resultados das séries criadas, em geral, foram superiores a série TRANSACTIONS, comprovando que há relevância na aplicação dos conhecimentos teóricos sobre o assunto de fraudes em criptomoedas para o desenvolvimento de séries temporais mais elaboradas, para serem dados de entrada em modelos de classificação baseados em RNA.

Mais especificamente, como dado relevante, a série temporal que apresentou uma melhor média aritmética de desempenho foi a de NEWUSER (85,8%), de acordo com a Tabela5.2.

Além disso, a série TRANSACTIONS apresentou um comportamento diferente das outras séries criadas: quanto maior a janela de tempo, menor ficou o desempenho.

Tal fato pode ser explicado ao analisar os gráficos das séries de 20, 40 e 60 dias, onde a diferença entre a média do total de transações de criptomoedas fraudulentas e não fraudulentas fica mais acentuada em 20 dias e vai diminuindo, ao chegar em 60 dias, como pode descrito nas Figuras 5.4a, 5.4b e 5.4c. Desta forma, quanto maior a diferença entre as séries fraudulentas e não fraudulentas, maior a capacidade, a priori, de a RNA identificar padrões de reconhecimento, a fim de classificá-las.

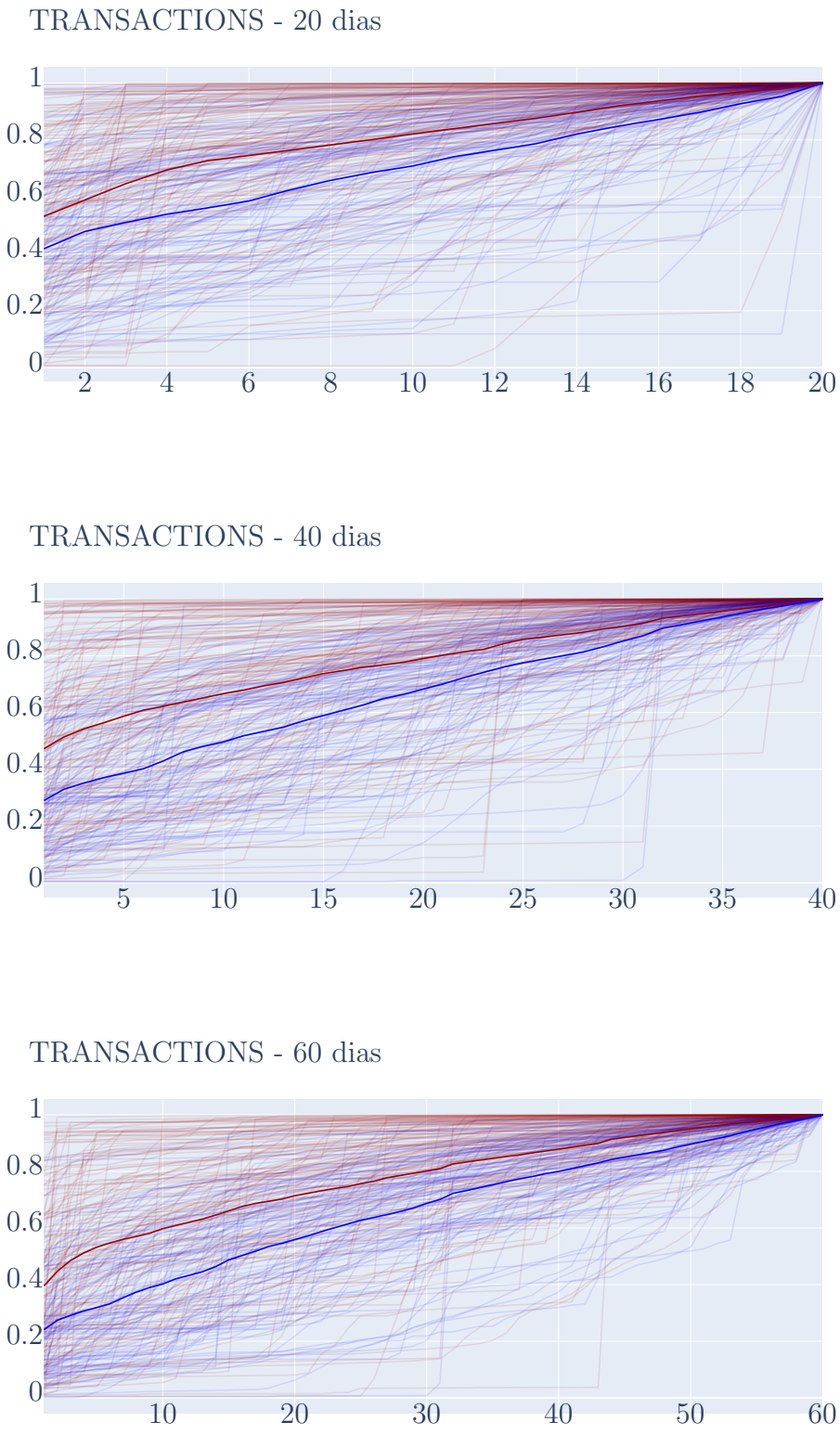


FIGURA 5.4 – Comparação entre as médias do número de transações ao longo do tempo (20, 40 e 60 dias).

TABELA 5.3 – Análise comparativa entre o tamanho de amostra de tempo e a média aritmética dos resultados

Janela de Tempo	20 dias	40 dias	60 dias
Média Aritmética *Recall*	83,6%	84,4%	83,2%

5.2 Comparação pelos tamanhos das janelas de tempo

A seguir, foi realizada uma comparação entre os tamanhos das janelas de tempo (20, 40 e 60 dias), com o objetivo de se verificar qual a melhor janela a se trabalhar.

A Figura 5.5, apresenta os resultados dos testes por janela de tempo. Pode-se observar que, no geral, há um pequeno aumento de desempenho, por ocasião do aumento do tamanho da amostra de tempo, com exceção do modelo LSTM-MLP, o qual apresentou uma queda brusca de desempenho, por ocasião do treinamento com a série de 60 dias. Será abordado este assunto com mais detalhes na próxima seção.

Séries / Tempo	**20 dias**			**40 dias**			**60 dias**		
	MLP	CNN-MLP	LSTM-MLP	MLP	CNN-MLP	LSTM-MLP	MLP	CNN-MLP	LSTM-MLP
NEWHOLDER	83%	86%	83%	76%	82%	91%	86%	81%	77%
NEWUSER	83%	83%	91%	86%	88%	88%	88%	90%	75%
BIGBUYER	86%	86%	80%	86%	86%	83%	88%	79%	75%
GAS/GASLIMIT	84%	75%	-	81%	81%	-	88%	88%	-

FIGURA 5.5 – Análise comparativa de métrica *Recall* pelo tamanho da amostra de dias

A Tabela5.3 descreve as médias aritméticas de desempenho dos testes, por janela de tempo. Se o modelo LSTM-MLP não apresentasse uma queda brusca de rendimento, a média das séries de 60 dias seria maior que as outras.

5.3 Comparação por modelo de RNA

E, por último, foi realizada uma análise comparativa entre os modelos de RNA abordados neste estudo (MLP, CNN-MLP e LSTM-MLP), com o objetivo de observar o comportamento de cada modelo, por ocasião da classificação de séries temporais. Os seguintes resultados foram coletados, segundo a Figura 5.6 e suas observações estão na Tabela5.4.

A queda de desempenho do modelo LSTM-MLP pode ser explicada pela pela análise da matriz de confusão, a partir dos resultados do modelo LSTM-MLP, utilizando-se as séries de 40 e 60 dias, como nas figuras 5.7a e 5.7b.

Tal diferença de resultado pode ser explicada a partir da análise de quatro criptomoedas erradamente preditas na série de 60 dias. Duas delas (grupo 1) também foram

Séries / RNA	MLP			CNN-MLP			LSTM-MLP		
	20 dias	40 dias	60 dias	20 dias	40 dias	60 dias	20 dias	40 dias	60 dias
NEWHOLDER	83%	76%	86%	86%	82%	81%	83%	91%	77%
NEWUSER	83%	86%	88%	83%	88%	90%	91%	88%	75%
BIGBUYER	86%	86%	88%	86%	86%	79%	80%	83%	75%
GAS/GASLIMIT	84%	81%	88%	75%	81%	88%	-	-	-
TRANSACTIONS	81%	76%	72%	88%	82%	68%	88%	85%	72%

FIGURA 5.6 – Análise comparativa de métrica *Recall* pelos modelos de RNA

TABELA 5.4 – Observações dos desempenhos dos modelos de RNA

Modelos	Observações
MLP	O modelo MLP se mostrou satisfatório para reconhecimento de padrões em séries temporais. Como fora citado, já é aplicado para resolução de problemas semelhantes, como reconhecimento de atividade humana. Em geral, o desempenho aumentou, na medida que a janela de tempo fora aumentada.
CNN-MLP	O modelo CNN-MLP obteve um rendimento parecido com o MLP. Embora o modelo convolução tenha sido projetado para reconhecimento de imagens, ele se mostrou muito aplicável para reconhecimento de padrões em séries temporais. Diferente do MLP, não apresentou um padrão uniforme de desempenho em relação às janelas de tempo abordadas. Apesar disso, manteve seu desempenho semelhante ao anterior.
LSTM-MLP	Apresentou um ótimo desempenho para as séries de 20 e 40 dias. Por outro lado, o modelo LSTM-MLP, apesar de ter sido idealizado para guardar memória ao longo do tempo, não apresentou bons resultados para reconhecimento de padrões ao ser aplicado na maior série abordada, 60 dias.

erradamente preditas na série de 40 dias e duas (grupo 2) foram corretamente preditas na mesma série. Foi elaborado um gráfico comparativo entre os dois grupos, conforme as Figuras 5.8a e 5.8b.

Pode ser observado que as criptomoedas do grupo 2 tiveram um pico entre 40 e 60 dias, ao passo que, no grupo 1, não tiveram. Sendo assim, é um indício de que o desempenho da Rede LSTM-MLP tenha caído, por ocasião da série de 60 dias.

No entanto, se mostrou bastante aplicável para séries curtas, obtendo-se dois resultados bastante expressivos de 91%, mostrando-se como uma excelente opção para análise de séries de 20 e 40 dias.

A média aritmética dos resultados de *Recall*, segundo os modelos de RNA abordados neste estudo, podem ser visualizados na Figura 5.9.

(a) **NEWUSER - LSTM - 40 dias**

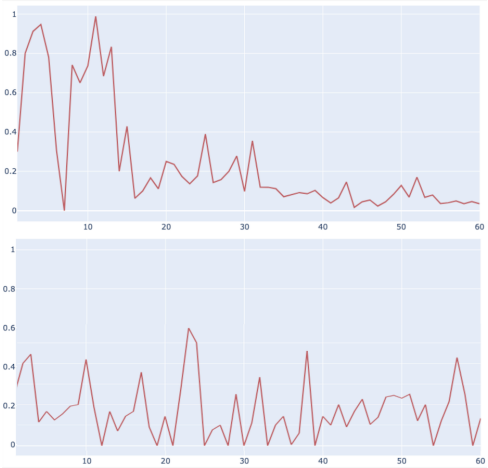
		Atualidade	
		0	1
Predição	0	TN = 09	FN = 05
	1	FP = 21	TP = 37

(b) **NEWUSER - LSTM - 60 dias**

		Atualidade	
		0	1
Predição	0	TN = 12	FN = 11
	1	FP = 15	TP = 35

FIGURA 5.7 – Matrizes de confusão originadas a partir dos resultados da Rede LSTM-MLP, aplicada às séries NEWUSERS de 40 e 60 dias, respectivamente.

(a) **Grupo 1**



(b) **Grupo 2**

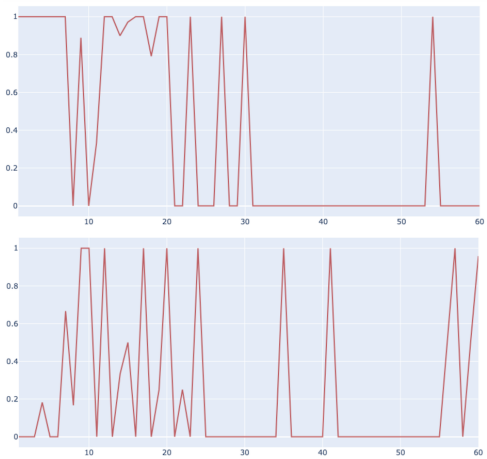


FIGURA 5.8 – Gráfico da série NEWUSER ao longo de 60 dias das criptomoedas do grupo 1 e 2.

A Tabela 5.5 descreve a média aritmética dos desempenhos dos três modelos de RNA apresentados neste estudo. Como pode ser observado, os três modelos atingiram médias semelhantes. O modelo MLP se mostrou mais uniforme quanto aos números. O modelo CNN-MLP apresentou menos uniformidade que o anterior. Todavia, obteve resultados expressivos (como de 88% e 90%) em algumas determinadas condições. E, por último, o modelo LSTM-MLP se mostrou como a melhor opção deste estudo para reconhecimento da série de curta duração (20 dias).

5.4 Síntese dos Resultados

Primeiramente, foram elencados os resultados das séries temporais criadas a partir dos Trabalhos Relacionados. Verificou-se que a série com maior desempenho foi a de NEWU-

RNA / Tempo	20 dias	40 dias	60 dias
MLP	83%	81%	84%
CNN-MLP	84%	84%	81%
LSTM-MLP	86%	87%	75%

FIGURA 5.9 – Análise comparativa da média aritmética dos resultados de *Recall* pelos modelos de RNA, ao longo das janelas de tempo

TABELA 5.5 – Análise Comparativa entre o tipo de modelo utilizado e as suas respectivas médias aritméticas de desempenho

Modelo de RNA	MLP	CNN-MLP	LSTM-MLP
Média Aritmética *Recall*	82,93%	82,87%	82,33%

SER, inspirada no trabalho (XU; LIVSHITS, 2019). Este estudo foi de suma importância pois, além de inspirar a construção desta série, inspirou outra série, a de NEWHOLDER.

A Série de BIGBUYER, inspirada no estudo de Milne, envolvendo Criptomoedas sob a perspectiva da Escola Austríaca (MILNE, 2018), convergiu para a conclusão desta pesquisa, a qual aborda que a descentralização das reservas de ativos é vista como um indicador de ausência de esquemas fraudulentos e, portanto, torna-se um atrativo para que os usuários venham a aderir a uma criptomoeda.

Ainda na análise dos resultados, de acordo com as séries temporais, realizou-se uma comparação entre as que foram inspiradas em Trabalhos Relacionados e a que não foi. Concluiu-se, por meio da análise dos resultados, que a influência dos Trabalhos Relacionados para a confecção das quatro primeiras séries contribuíram para o aumento do desempenho dos modelos de RNA, para detectar fraudes.

Em segundo lugar, efetuou-se uma comparação pelas janelas de tempo (20, 40 e 60 dias). A janela de tempo com maior desempenho, na média, foi a de 40 dias. Se o modelo LSTM-MLP não tivesse apresentado uma queda de desempenho, provavelmente por *outliers* na janela de tempo de 60 dias, a média dos resultados desta janela seria maior.

Por último, foi realizada uma comparação dos resultados pelos desempenhos dos modelos de RNA. Calculou-se a média aritmética dos desempenhos de cada modelo e o modelo MLP apresentou o maior desempenho. Assim como no caso das janelas de tempo, se o modelo LSTM-MLP não tivesse apresentado uma queda de desempenho, por ocasião da janela de 60 dias, a média dos resultados deste modelo seria maior.

O próximo Capítulo 6 sintetiza esta pesquisa, em termos de contribuições, trabalhos

futuros e limitações.

6 Conclusão

Esta pesquisa teve como principal objetivo o desenvolvimento de um método para a detecção de fraudes em criptomoedas baseadas na rede Ethereum e advindas de atividades de ICO.

A motivação do trabalho, em capturar fraudes, teve base na popularidade das atividades de ICO do final do ano de 2018. Como suas transações aconteceram ao longo do tempo, entre vários usuários distintos ao longo da rede Ethereum, a detecção de fraudes seria de suma importância para que os usuários pudessem retirar seus fundos o mais depressa possível, sem serem lesados por malfeitores.

A natureza das transações que acontecem em sequência, ao longo do tempo, assim como o preço de ações no mercado financeiro, justificou o uso de ferramentas de predição em séries temporais para resolver este tipo de problema. Desta forma, os modelos criados para a detecção de fraudes foram desenvolvidos utilizando-se o estado da arte em termos de redes neurais para a classificação de séries temporais.

Adicionalmente, durante este trabalho, foi realizado um estudo preliminar de detecção dos padrões, os quais poderiam servir de base para a confecção das séries temporais. Foram criadas cinco séries, quatro inspiradas nos Trabalhos Relacionados (NEWHOLDER, NEWUSER, BIGBUYER e GASGASLIMIT) e uma com base na quantidade de transações ao longo do tempo (TRANSACTIONS). Para cada série, foram projetadas três janelas de tempo, tornando-se um total de 15 séries a serem entradas nos modelos de RNA.

Finalmente, as séries temporais passaram pelos três modelos de RNA projetados para a classificação (MLP, CNN-MLP e LSTM-MLP). Ao todo, foram realizados 45 experimentos, de modo a verificar qual seria, em termos de desempenho (*Recall*, a melhor série criada, a melhor janela de tempo e o melhor modelo de RNA projetado.

6.1 Contribuições deste trabalho

De modo a descrever, quantitativamente, as principais contribuições do trabalho, esta seção resume os principais pontos.

Em primeiro lugar, verificou-se que modelos precisos podem ser criados com RNAs para classificação de séries temporais, geradas a partir do histórico de transações das criptomoedas. As séries temporais criadas ao longo deste trabalho, baseadas em hipóteses apresentadas de seleção dos melhores parâmetros, se mostraram mais eficientes para detecção de fraudes do que uma série simples, normalizada, atinente ao número de transações por dia. Além disto, os modelos classificatórios apresentados neste estudo se mostraram com valores de (*Recall*) superiores aos Trabalhos Relacionados presentes nesta pesquisa, (CHEN *et al.*, 2018) que obteve 81% e (CHEN *et al.*, 2019), com desempenho obtido de 69%.

Adicionalmente, este trabalho mostrou-se adaptável, não somente detectando um tipo de fraude, mas também generalizando para fraudes de Esquemas de Saída, *Pump and Dumps* e Esquemas Ponzi, conforme descritas na seção de Fundamentação Teórica pois, nos vários sub-conjuntos de criptomoedas fraudulentas estudados, encontram-se incluídos diversos exemplares destes esquemas.

O resultado da Série BIGBUYER, por exemplo, convergiu com o artigo Criptomoedas sob Perspectiva da Escola Austríaca (MILNE, 2018), o qual aborda que a descentralização de reserva monetária é um atrativo para a adesão de investidores em criptomoedas.

Também foi realizado um estudo de Análise Exploratória de Dados (EDA), que auxiliou a selecionar características das séries de maior relevância, um aspecto pouco explorado na literatura. Por exemplo, informações importantes foram extraídas como a diferença entre a data da primeira transação da criptomoeda e a data de sua entrada no mercado, tipo de conta (contrato, casa de câmbio ou outro tipo) do maior detentor de títulos e a média de transações de novos usuários criados na rede Ethereum. Este último parâmetro serviu de base para o desenvolvimento da Série NEWUSER, que obteve o melhor desempenho dentre os modelos de rede neural testados.

Do ponto de vista técnico, foram analisados, ao todo, 238 *dataset* de criptomoedas, sendo 136 fraudulentas e 102 não-fraudulentas. O tamanho de cada *dataset* de transações, em formato CSV, variou desde 450KB até 450MB, aproximadamente. O tempo de coleta destas bases de dados aproximou-se de 1 mês. Após um processo de limpeza dos dados, foram extraídos dos arquivos somente os valores a serem usados como entradas nos modelos de classificação. Deste modo, ficou otimizado o tempo de treinamento das RNAs. Cada RNA, a partir dos arquivos de texto, demorou, em média, 30 segundos para pré-processamento e treinamento, sendo que o modelo LSTM-MLP foi o que mais demorou a convergir, levando em média 2 minutos completos para treinamento e predição. Foi

utilizado no processamento dos dados um computador Intel 64 bits com 8 cores e 64GB de RAM, disco SSD e 2 placas GPU NVidia 1080i.

Os resultados foram encorajadores, pois foi possível detectar fraudes nas ICOs com *Recall* de 91% para amostras de tempo de 20 dias após o lançamento da criptomoeda no mercado. Portanto, tal modelo apresentou um horizonte razoável (20 dias) para o pequeno investidor, a fim de servir como subsídio para a a entrada em uma atividade de ICO.

Com relação ao desempenho das redes neurais artificiais utilizadas, é possível resumir as seguintes observações neste estudo. A arquitetura CNN, normalmente utilizada para resolver problemas que envolvam reconhecimento de imagem (base do modelo CNN-MLP), se mostrou aplicável à classificação de séries temporais, chegando, neste estudo, a um valor de até 90% de *Recall* (modelo CNN-MLP). Além disto, embora as séries desenvolvidas nesta pesquisa não tenham apresentado clara tendência, sazonalidade e ciclo, o modelo LSTM-MLP, projetado para o aprendizado baseado em observações genéricas passadas, conseguiu efetuar o treinamento, chegando a convergir com valor máximo de 91% de *Recall*.

6.2 Trabalhos Futuros

Nem todos os aspectos abordados neste estudo foram aprofundados. Por exemplo, investigar um parâmetro para uma série temporal que descreva a porcentagem de títulos que o maior detentor de títulos possui pode vir a ser uma sugestão de trabalho futuro promissora. Porém, seria necessário realizar varreduras amplas nos códigos dos contratos inteligentes referentes a cada criptomoeda estudada.

Com relação aos modelos de séries temporais, uma pergunta pendente que carece de maior investigação é se a queda de desempenho do modelo LTSM-MLP nas sub-séries de 60 dias representa algo significativo e específico de situações que envolvam criptomoedas ou genéricos. Ainda neste modelo, mais dados precisam ser processados para entender a razão da impossibilidade de treinar a Série Temporal GAS/GASLIMIT. Novamente a pergunta é sobre a generalidade deste resultado. Ainda, em trabalhos futuros, é desejável se aplicar as séries temporais criadas em modelos híbridos (CNN-LSTM-MLP e CONV-LSTM). Modelos híbridos estes abordados anteriormente na Fundamentação Teórica, de modo, a estudar se eles são aplicáveis para a resolução deste tipo de problema.

Adicionalmente, pode-se, futuramente, desenvolver um serviço de identificação de fraudes em criptomoedas, baseado neste estudo, que auxilie os futuros investidores a entrar no mercado de criptomoedas com mais segurança, bem como maximizar os seus lucros e diminuir perdas.

Finalmente, uma frente de estudo poderia ser aberta para o desenvolvimento de modelos de predição em tempo real, pois as transações de ICOs de criptomoedas são atualizadas constantemente ao longo do tempo.

6.3 Limitações do Trabalho

Apesar de alcançar todos os objetivos anteriormente citados e realizar contribuições para a área, entretanto, algumas limitações foram observadas, ao longo deste estudo.

Considerando-se o problema da limpeza dos dados, por ocasião da coleta do dia em que a criptomoeda entrou no mercado, foram observadas várias inconsistências de datas nas plataformas que proveem informações sobre atividades de ICOs. Portanto, ficou estabelecido como limitação o critério de usar o primeiro dia em que a criptomoeda superou 400 transações como Data de Entrada no Mercado.

Ao utilizar as plataformas de verificação de transações na rede Ethereum, como a API da plataforma Etherscan, esta limita a coleta a somente 5 requisições por segundo. Portanto, a coleta e montagem de séries como BIGBUYER, passo em que é verificado se o maior comprador é ou não uma casa de câmbio, teve longo tempo de processamento. Esses dados brutos também possuem múltiplas dimensões, o que onera o processamento. Portanto, foram criados arquivos específicos somente com os dados filtrados e considerados como relevantes. Dependendo da maneira com que um processamento de tempo real opere, isto pode ser considerado como um limitador.

A série BIGBUYER foi inicialmente planejada para ser BIGHOLDER, maior detentor de títulos, ao invés de maior comprador. Contudo, verificou-se que haviam criptomoedas que apresentavam uma diferença entre a disposição de maiores detentores de títulos e a disposição de maiores compradores de títulos. É possível que isto se deva ao contrato inteligente inicial da respectiva atividade de ICO da criptomoeda, que estaria gerando *tokens* para determinados usuários. Entretanto, tudo indica que mais estudos são necessários para ultrapassar também esta limitação.

Referências

ABDELHAMID, M.; HASSAN, G. Blockchain and smart contracts. In: ACM. **Proceedings of the 2019 8th International Conference on Software and Information Engineering**. [S.l.], 2019. p. 91–95.

ADHAMI, S.; GIUDICI, G.; MARTINAZZI, S. Why do businesses go crypto? an empirical analysis of initial coin offerings. **Journal of Economics and Business**, Elsevier, v. 100, p. 64–75, 2018.

ANGUITA, D.; GHIO, A.; ONETO, L.; PARRA, X.; REYES-ORTIZ, J. L. A public domain dataset for human activity recognition using smartphones. In: **ESANN**. [S.l.: s.n.], 2013. v. 3, p. 3.

ANTONOPOULOS, A. M.; WOOD, G. **Mastering ethereum: building smart contracts and dapps**. [S.l.]: O'Reilly Media, 2018.

BANOS, O.; GALVEZ, J.-M.; DAMAS, M.; POMARES, H.; ROJAS, I. Window size impact in human activity recognition. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 14, n. 4, p. 6474–6499, 2014.

BARTOLETTI, M.; CARTA, S.; CIMOLI, T.; SAIA, R. Dissecting ponzi schemes on ethereum: identification, analysis, and impact. **Future Generation Computer Systems**, Elsevier, v. 102, p. 259–277, 2020.

BAUM, S. C. Cryptocurrency fraud: A look into the frontier of fraud. Georgia Southern University, 2018.

BECK, R.; MÜLLER-BLOCH, C. Blockchain as radical innovation: a framework for engaging with distributed ledgers as incumbent organization. 50th Hawaii International Conference on System Sciences, 2017.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C. **Time series analysis: forecasting and control**. [S.l.]: John Wiley & Sons, 2011.

BRAGA, A. d. P. **Redes neurais artificiais: teoria e aplicações**. [S.l.]: Livros Técnicos e Científicos, 2000.

BROWNLEE, J. **Deep learning for time series forecasting: Predict the future with MLPs, CNNs and LSTMs in Python**. [S.l.]: Machine Learning Mastery, 2018.

BUTERIN, V. *et al.* Ethereum white paper: a next generation smart contract & decentralized application platform. 2014.

CASINO, F.; DASAKLIS, T. K.; PATSAKIS, C. A systematic literature review of blockchain-based applications: current status, classification and open issues. **Telematics and Informatics**, Elsevier, 2018.

CERQUEIRA, E. O. d.; ANDRADE, J. C. d.; POPPI, R. J.; MELLO, C. Redes neurais e suas aplicações em calibração multivariada. **Química Nova**, SciELO Brasil, v. 24, n. 6, p. 864–873, 2001.

CHEN, W.; ZHENG, Z.; CUI, J.; NGAI, E.; ZHENG, P.; ZHOU, Y. Detecting ponzi schemes on ethereum: Towards healthier blockchain technology. In: **Proceedings of the 2018 World Wide Web Conference**. [S.l.: s.n.], 2018. p. 1409–1418.

CHEN, W.; ZHENG, Z.; NGAI, E. C.-H.; ZHENG, P.; ZHOU, Y. Exploiting blockchain data to detect smart ponzi schemes on ethereum. **IEEE Access**, IEEE, v. 7, p. 37575–37586, 2019.

CHOD, J.; LYANDRES, E. A theory of icos: Diversification, agency, and information asymmetry. **Agency, and Information Asymmetry**, 2019.

CONLEY, J. P. Blockchain and the economics of crypto-tokens and initial coin offerings (no. 17-00008). **Nashville: Vanderbilt University**, 2017.

DANNEN, C. **Introducing Ethereum and Solidity**. [S.l.]: Springer, 2019.

FARELL, R. An analysis of the cryptocurrency industry. University of Pennsylvania, 2015.

FRIZZO-BARKER, J.; CHOW-WHITE, P. A.; ADAMS, P. R.; MENTANKO, J.; HA, D.; GREEN, S. Blockchain as a disruptive technology for business: A systematic review. **International Journal of Information Management**, Elsevier, v. 51, p. 102029, 2020.

GIUDICI, G.; MILNE, A.; VINOGRADOV, D. Cryptocurrencies: market analysis and perspectives. **Journal of Industrial and Business Economics**, Springer, v. 47, n. 1, p. 1–18, 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. `http://www.deeplearningbook.org`.

GREVE, F. G.; SAMPAIO, L. S.; ABIJAUDE, J. A.; COUTINHO, A. C.; VALCY, Í. V.; QUEIROZ, S. Q. Blockchain e a revolução do consenso sob demanda. **Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)-Minicursos**, 2018.

HAHN, C.; WONS, A. **Initial Coin Offering (ICO): Unternehmensfinanzierung auf Basis der Blockchain-Technologie**. [S.l.]: Springer, 2018.

HARTMANN, F.; WANG, X.; LUNESU, M. I. Evaluation of initial cryptoasset offerings: the state of the practice. In: IEEE. **2018 International Workshop on Blockchain Oriented Software Engineering (IWBOSE)**. [S.l.], 2018. p. 33–39.

HAYKIN, S. S. *et al.* **Neural networks and learning machines/Simon Haykin.** [S.l.]: New York: Prentice Hall, 2009.

HUANG, W.; MEOLI, M.; VISMARA, S. The geography of initial coin offerings. **Small Business Economics**, Springer, v. 55, n. 1, p. 77–102, 2020.

KEOGH, E.; CHU, S.; HART, D.; PAZZANI, M. Segmenting time series: A survey and novel approach. In: **Data mining in time series databases**. [S.l.]: World Scientific, 2004. p. 1–21.

KHER, R.; TERJESEN, S.; LIU, C. Blockchain, bitcoin, and icos: a review and research agenda. **Small Business Economics**, Springer, p. 1–22, 2020.

KOSTADINOV, S. **Recurrent Neural Networks with Python Quick Start Guide: Sequential Learning and Language Modeling with TensorFlow**. [S.l.]: Packt Publishing Ltd, 2018.

LANSKY, J. Cryptocurrency survival analysis. **The Journal of Alternative Investments**, Institutional Investor Journals Umbrella, v. 22, n. 3, p. 55–64, 2019.

LAURENCE, T. **Blockchain for dummies**. [S.l.]: John Wiley & Sons, 2019.

LECUN, Y.; BENGIO, Y. *et al.* Convolutional networks for images, speech, and time series. **The handbook of brain theory and neural networks**, v. 3361, n. 10, p. 1995, 1995.

MEHROTRA, K.; MOHAN, C. K.; RANKA, S. **Elements of artificial neural networks**. [S.l.]: MIT press, 1997.

MILLS, T. C. **Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting**. [S.l.]: Academic Press, 2019.

MILNE, A. Cryptocurrencies from an austrian perspective. In: **Banking and Monetary Policy from the Perspective of Austrian Economics**. [S.l.]: Springer, 2018. p. 223–257.

MUKHOPADHYAY, M. **Ethereum Smart Contract Development: Build blockchain-based decentralized applications using solidity**. [S.l.]: Packt Publishing Ltd, 2018.

MURAD, A.; PYUN, J.-Y. Deep recurrent neural networks for human activity recognition. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 17, n. 11, p. 2556, 2017.

NAKAMOTO, S. **Bitcoin: A peer-to-peer electronic cash system**. [S.l.], 2019.

NARAYANAN, A.; BONNEAU, J.; FELTEN, E.; MILLER, A.; GOLDFEDER, S. Bitcoin and cryptocurrency technologies. **Curso elaborado pela Princeton University Press**, 2019.

NIELSEN, M.; BENGIO, Y.; COUVILLE, A. Deep learning. **Retrieved from http://neuralnetworksanddeeplearning**, 2017.

OLIVA, G. A.; HASSAN, A. E.; JIANG, Z. M. J. An exploratory study of smart contracts in the ethereum blockchain platform. **Empirical Software Engineering**, Springer, p. 1–41, 2020.

ORDÓÑEZ, F. J.; ROGGEN, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 16, n. 1, p. 115, 2016.

PAL, A.; PRAKASH, P. **Practical Time Series Analysis: Master Time Series Data Processing, Visualization, and Modeling using Python**. [S.l.]: Packt Publishing Ltd, 2017.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. **Bioinfo Publications**, 2011.

QUEST, M. The ico approach: A beginner's guide to understanding cryptocurrency ico. CreateSpace **Independent Publishing Platform**, 2018.

ROSEBROCK, A. **Deep Learning for Computer Vision with Python: Starter Bundle**. [S.l.]: PyImageSearch, 2017.

SOLOLON. **Ethereum for dummies**. [S.l.]: John Wiley & Sons, 2019.

SZABO, N. Formalizing and securing relationships on public networks. **First monday**, 1997.

ULRICH, F. **Bitcoin: a moeda na era digital**. [S.l.]: LVM Editora, 2017.

WANG, J.; CHEN, Y.; HAO, S.; PENG, X.; HU, L. Deep learning for sensor-based activity recognition: A survey. **Pattern Recognition Letters**, Elsevier, v. 119, p. 3–11, 2019.

XU, J.; LIVSHITS, B. The anatomy of a cryptocurrency pump-and-dump scheme. In: **28th USENIX Security Symposium (USENIX Security 19)**. [S.l.: s.n.], 2019. p. 1609–1625.

XU, X.; PAUTASSO, C.; ZHU, L.; GRAMOLI, V.; PONOMAREV, A.; TRAN, A. B.; CHEN, S. The blockchain as a software connector. In: IEEE. **2016 13th Working IEEE/IFIP Conference on Software Architecture (WICSA)**. [S.l.], 2016. p. 182–191.

YANG, J.; NGUYEN, M. N.; SAN, P. P.; LI, X.; KRISHNASWAMY, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In: CITESEER. **Ijcai**. [S.l.], 2015. v. 15, p. 3995–4001.

ZENG, M.; NGUYEN, L. T.; YU, B.; MENGSHOEL, O. J.; ZHU, J.; WU, P.; ZHANG, J. Convolutional neural networks for human activity recognition using mobile sensors. In: IEEE. **6th International Conference on Mobile Computing, Applications and Services**. [S.l.], 2014. p. 197–205.

ZHAO, J. L.; FAN, S.; YAN, J. **Overview of business innovations and research opportunities in blockchain and introduction to the special issue**. [S.l.]: SpringerOpen, 2016.

ZURADA, J. **Introduction to artificial neural system**. [S.l.]: West Pub. Co., 1992.

Apêndice A - Glossário de Termos Técnicos

Arquitetura de RNA	Estrutura de uma RNA. Neste trabalho, foram utilizadas as MLP, CNN e LSTM.
Análise	Exame detalhado de cada parte que compõe um todo, buscando compreender tudo aquilo que o caracteriza.
API	Em português, Interface de Programação de Aplicativos, baseada na web, que representa um conjunto de rotinas e padrões de programação para acesso a aplicativos de software ou plataformas.
ARIMA	Média Móvel Integrada Autoregressiva (Autoregressive Integrated Moving Average). É uma técnica comumente usada para fazer análise de predição.
Avaliação	Apreciação quantitativa e/ou qualitativa sobre dados relevantes do processo.
Atividade de ICO	Conjunto das ações realizadas durante as 3 fases de Oferta Inicial de Moedas (Initial Coin Offering - ICO): Pré-ICO, Lançamento e Pós-ICO.
Criptomoeda Fraudulenta	Criptomoeda cuja organização fundadora esteja envolvida em alguma atividade fraudulenta.
Eficácia	Capacidade de cumprir os objetivos pretendidos.
Eficiência	Poder de realizar algo, convenientemente, despendendo de um mínimo de esforço, tempo e outros recursos ou competências.

Experimento	Pesquisa planejada para obter novos fatos, para obter ou confirmar resultados, tendo por objetivo tomar decisões. No contexto deste trabalho, foram realizados 45 experimentos, cuja a quantidade foi baseada na combinação das variáveis: 5 séries temporais, 3 janelas de tempo e 3 modelos de RNA.
Hipótese	Neste trabalho, é uma intuição, baseada nos Trabalhos Relacionados. Não confundir com testes de hipóteses em Estatística.
Método	Conjunto de passos para se chegar a um objetivo, de forma repetível, obedecendo a um conjunto de regras.
Metodologia	Estudo do método.
Métrica de Desempenho	Medida quantitativa que indica o grau que um modelo possui determinado atributo. Neste trabalho, foi utilizado Recall.
Modelo de RNA	Representação de uma RNA, composto por uma ou mais arquiteturas de RNA. Neste trabalho, foram projetados 3 modelos de RNA: MLP, CNN-MLP e LSTM-MLP.
Passo de um Método	Parte de um método.
Processo	Conjunto sequencial e particular de ações com objetivo.
Rede Ethereum	É a Tecnologia Blockchain do projeto Ethereum.
Técnica	Maneira pela qual um ou mais passos de um método pode ser executado.
Tecnologia Blockchain	É a estrutura blockchain aplicada à realidade.
Token	Neste trabalho, é uma criptomoeda desenvolvida na rede Ethereum.
Validação	É o ato de tornar ou declarar algo como válido.

Verificação	É o ato de examinar se algo é o que deve ser ou o que se declarou ser.

FOLHA DE REGISTRO DO DOCUMENTO

[1.] CLASSIFICAÇÃO/TIPO	[2.] DATA	[3.] REGISTRO N°	[4.] N° DE PÁGINAS
DM	05 de abril de 2021	DCTA/ITA/DM-016/2021	119

[5.] TÍTULO E SUBTÍTULO:

Detecção de fraudes em criptomoedas utilizando métodos de classificação de séries temporais baseados em redes neurais.

[6.] AUTOR(ES):

Luiz Alfredo Zenon da Mata Caffé

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

[8.] PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Mises; Criptomoeda; Oferta Inicial de Moedas; Blockchain; Séries Temporais; Redes Neurais

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Análise de séries temporais; Redes neurais; Protocolo de confiança; Redes de comunicação; Inteligência artificial; Computação.

[10.] APRESENTAÇÃO: **(X) Nacional ( ) Internacional**

ITA, São José dos Campos. Curso de Mestrado. Programa de Pós-Graduação em Engenharia Eletrônica e Computação. Área de Informática. Orientador: Prof. Cesar Augusto Cavalheiro Marcondes. Defesa em 04/03/2021. Publicada em 2021.

[11.] RESUMO:

Este trabalho apresenta um método para a detecção de fraudes em criptomoedas, originadas a partir de uma Oferta Inicial de Moedas (Initial Coin Offering - ICO). Para isto, foram utilizados modelos preditivos, baseados em redes neurais, para a classificação de Séries Temporais, geradas a partir das tabelas de fluxo de transações na rede Ethereum. A primeira atividade de ICO foi executada em 2013 e alcançou o seu auge no primeiro semestre de 2018, com movimentações entre 7 e 12 bilhões de USD em todo o mundo. Todavia, estima-se que 78% das atividades de ICO são fraudulentas. Baseadas no comportamento de criptomoedas fraudulentas e não fraudulentas, bem como nas tabelas de transações das criptomoedas coletadas, foram desenvolvidas 5 séries temporais normalizadas, que deram entrada nos modelos de Redes Neurais Artificiais (RNA) dos tipos Multi Layer Perceptron (MLP), Convolution Neural Network - Multi Layer Perceptron (CNN-MLP) e Long Short Term Memory - Multi Layer Perceptron (LSTM-MLP) projetados para classificação. Ao final da pesquisa, foi obtido um valor de (Recall) de até 91% em alguns casos.

[12.] GRAU DE SIGILO:

(X) OSTENSIVO ( ) RESERVADO ( ) SECRETO