UNIVERSIDADE DE BRASÍLIA
Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas
Departamento de Economia
Programa de Pós-Graduação em Economia

HENRIQUE YASSUYUKI TSUBOI

IDENTIFICAÇÃO DE ATIVIDADES SUSPEITAS DE LAVAGEM DE DINHEIRO: ABORDAGEM COM APRENDIZAGEM DE MÁQUINA EM CARTEIRAS NA REDE ETHEREUM

Brasília/DF
2023

HENRIQUE YASSUYUKI TSUBOI

IDENTIFICAÇÃO DE ATIVIDADES SUSPEITAS DE LAVAGEM DE DINHEIRO: ABORDAGEM COM APRENDIZAGEM DE MÁQUINA EM CARTEIRAS NA REDE ETHEREUM

Dissertação apresentada ao Programa de Pós-Graduação em Economia, do Departamento de Economia da Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas da Universidade de Brasília, como requisito parcial à obtenção do título de Mestre em Economia

Orientador: Prof. Dr. Rafael Terra de Menezes
Coorientador: Prof. Dr. Rafael Sousa Lima

Brasília/DF
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

TT882i Tsuboi, Henrique Yassuyuki
IDENTIFICAÇÃO DE ATIVIDADES SUSPEITAS DE LAVAGEM DE DINHEIRO: ABORDAGEM COM APRENDIZAGEM DE MÁQUINA EM CARTEIRAS NA REDE ETHEREUM / Henrique Yassuyuki Tsuboi; orientador Rafael Terra de Menezes; co-orientador Rafael Sousa Lima. -- Brasília, 2023.
51 p.

Dissertação (Mestrado Profissional em Economia - Gestão Econômica de Finanças Públicas) -- Universidade de Brasília, 2023.

1. Ethereum. 2. Criptoativos. 3. Lavagem de Dinheiro. 4. Aprendizado de Máquina. 5. LightGBM. I. Menezes, Rafael Terra de, orient. II. Lima, Rafael Sousa, co-orient. III. Título.

HENRIQUE YASSUYUKI TSUBOI

IDENTIFICAÇÃO DE ATIVIDADES SUSPEITAS DE LAVAGEM DE DINHEIRO: ABORDAGEM COM APRENDIZAGEM DE MÁQUINA EM CARTEIRAS NA REDE ETHEREUM

Dissertação apresentada ao Programa de Pós-Graduação em Economia, do Departamento de Economia da Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas da Universidade de Brasília, como requisito parcial à obtenção do título de Mestre em Economia

BANCA EXAMINADORA:

Prof. Dr. Rafael Terra de Menezes

Departamento de Economia – UnB

Orientador

Prof.ª Dra. Ana Carolina Pereira Zoghbi

Departamento de Economia – UnB

Examinadora Interna

Prof. Dr. Kleber Vasconcelos de Oliveira

Banco Central do Brasil

Examinador Externo

Brasília/DF
2023

AGRADECIMENTOS

À Universidade de Brasília, expresso minha gratidão pela oportunidade de fazer parte deste programa de pós-graduação.

Aos organizadores deste Mestrado Profissional em Economia, que se esforçaram para a sua realização.

Agradeço a todos os professores da instituição, incluindo o meu orientador Prof. Dr. Rafael Terra de Menezes, pelo ensino de alta qualidade que recebi e a Prof. Dra. Ana Carolina Pereira Zoghbi pelo empenho, dedicação e profissionalismo.

Ao meu Coorientador, Prof. Dr. Rafael Sousa Lima, que sempre me incentivou e teve muita paciência comigo no decurso desse projeto. Seus ensinamentos e calma possibilitaram a finalização deste trabalho.

Ao meu grande amigo Suto, que me ajudou e me incentivou até a reta final deste trabalho. Muita gratidão.

Aos meus colegas de turma, que se empenharam para alcançar a conclusão deste mestrado, sobretudo à Flávia que na etapa final deste projeto me incentivou até a sua conclusão.

Aos meus amigos Silvia e Aurélio, pela constante torcida.

A todos vocês muito obrigado!!!

RESUMO

Esse trabalho tem como objetivo investigar em que medida é possível identificar endereços de criptomoedas suspeitos de envolvimento com a prática da lavagem de dinheiro. A motivação da pesquisa tem por base e inspiração o aumento da utilização mundial da tecnologia blockchain e de criptomoedas como o Ethereum, chamando atenção para a sua utilização por parte de criminosos em práticas ilícitas, como a lavagem de dinheiro. A finalidade desta pesquisa é classificar endereços de criptomoedas como lícitos e ilícitos, por meio do emprego de técnicas de aprendizado de máquina. Neste trabalho foi adotado o modelo de *machine learning* conhecido por LightGBM, em busca de sinalizações que pudessem remeter a atividades suspeitas de lavagem de dinheiro. Os resultados apontam que o atributo mais influenciador na identificação de uma carteira ilícita foi o "*Time_diff_between_ first_and_last_(Mins)*", que indica a diferença de tempo entre a primeira e a última atividade da carteira, indicando um período curto de "vida" do endereço considerado ilícito, o que pode sugerir seu uso em práticas como "*pooling accounts*", técnica de lavagem de dinheiro no qual é utilizada inúmeras transferências em contas de passagem, o que dificulta o rastreamento da origem dos valores. Acredita-se que o tema é atual e relevante, indo ao encontro de recentes inovações legislativas no mercado brasileiro de criptoativos. Ademais, a pesquisa se mostra relevante do ponto de vista prático e acadêmico, pois oferece percepções valiosas tanto para órgãos reguladores como para usuários, além de ajudar a desmistificar o uso das técnicas de máquina de aprendizagem em pesquisas acadêmicas no campo da ciência social.

Palavras-Chave: Ethereum; Criptoativos; Lavagem de Dinheiro; Aprendizado de Máquina; LightGBM.

ABSTRACT

This study aims to investigate to what extent it is possible to identify addresses of cryptocurrencies suspected of being involved in money laundering activities. The research motivation is based on the increasing worldwide use of blockchain technology and cryptocurrencies such as Ethereum, drawing attention to their utilization by criminals in illicit activities, such as money laundering. The purpose of this research is to classify cryptocurrency addresses as licit or ilicit through the use of machine learning techniques. In this work, the machine learning model known as LightGBM was adopted in search of indicators that could be associated with suspicious money laundering activities. The results indicate that the most influential attribute in identifying an illicit wallet is the "Time_diff_between_first_and_last_(Mins)", which indicates the time difference between the first and last activity of the wallet, suggesting a short lifespan of the considered illicit address, which may imply its use in practices such as “pooling accounts", a money laundering technique that involves multiple transfers through intermediary accounts, making it difficult to trace the origin of the funds. It is believed that the topic is current and relevant, aligning with recent legislative innovations in the Brazilian market of crypto assets. Furthermore, the research proves to be relevant from both a practical and academic perspective, as it offers valuable insights to regulatory bodies and users alike, while also helping to demystify the use of machine learning techniques in social science research.

Keywords: Ethereum; Crypto assets; Money Laundering; Machine Learning; LightGBM

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AML	*Anti-Money Laundering* (Anti Prevenção a Lavagem de Dinheiro)
DAO	*Decentralized Autonomous Organization*
dApps	*Decentralized applications*
DDoS	*Distributed Denial of Service* (Negação de Serviço Distribuída)
DEX	*Decentralized Exchanges*
ETH	Ether
FATF	*Financial Action Force*
KYC	*Know Your Custormer* (Conheça seu cliente)
LightGBM	*Light Gradient Boosting Machine*
NFT	*Non-fungible tokens*
ROC-AUC	*Receiver Operating Characteristic - Area Under the Curve*
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

A globalização e os desafios trazidos com o seu advento resultaram no aparecimento de diversas ideias e tecnologias, dentre elas a tecnologia *blockchain*, que foi globalmente conhecida no final da década de 2000, com a publicação do artigo *Bitcoin: A Peer-to-Peer Electronic Cash System* por Satoshi Nakamoto (2008). O artigo descrevia um sistema descentralizado e seguro de transações financeiras que utilizava criptografia para proteger a integridade dos registros de transações em um registro público compartilhado, conhecido como blockchain. O sistema foi implementado em 2009 com o lançamento da primeira criptomoeda, o Bitcoin. Desde então, a tecnologia blockchain evoluiu e foi aplicada em diversas áreas, incluindo finanças, saúde, logística, votação eletrônica, entre outras.

A tecnologia blockchain foi base para o surgimento de novos projetos, como o Ethereum, que foi criada por Vitalik Buterin em 2015 como uma forma de expandir as capacidades da tecnologia de blockchain do Bitcoin, além das transações de moeda digital. O Ethereum é uma plataforma de blockchain descentralizada que permite a criação, desenvolvimento e execução de contratos inteligentes e aplicativos descentralizados (dApps). Contratos inteligentes são programas autônomos que são executados em redes blockchain, possuem regras predefinidas e condições que são programadas em seu código, e uma vez que as condições são cumpridas, o contrato é executado automaticamente, sem a necessidade de intervenção humana, enquanto dApps são os aplicativos completos que utilizam esses contratos inteligentes como parte de sua infraestrutura para oferecer funcionalidades descentralizadas aos usuários, como mercado de tokens não fungíveis (NFTs), identidade digital, governança e muito mais. O Ethereum tem sua própria criptomoeda chamada Ether (ETH), que é usada para facilitar transações e incentivar desenvolvedores a construir e executar dApps na plataforma.

Uma das principais características e inovação do Ethereum, se comparado ao Bitcoin, é a capacidade de criar e executar contratos inteligentes, que são contratos autoexecutáveis com os termos do acordo diretamente escritos em código. Os contratos inteligentes podem ser usados para automatizar e aplicar os termos de vários tipos de acordos, como contratos financeiros, gerenciamento de cadeia de suprimentos e verificação de identidade.

Ethereum também é conhecido por sua capacidade de suportar organizações autônomas descentralizadas (DAOs), que são organizações que são executadas por meio de regras codificadas como programas de computador no blockchain do Ethereum e as decisões e a governança da organização são tomadas por meio de mecanismos de votação e consenso entre os participantes, que podem ser detentores de tokens ou partes interessadas no projeto, sem a

necessidade de uma autoridade ou gerenciamento centralizado. Em geral, o Ethereum fornece uma plataforma poderosa para desenvolvedores criarem aplicativos descentralizados e possibilitar uma nova geração de sistemas e organizações descentralizadas.

O crescimento da utilização e da aceitação do uso de criptoativos em transações financeiras despertou o interesse de criminosos, no potencial que esse novo mercado apresentava para a lavagem de dinheiro. A lavagem de dinheiro é uma prática ilegal, na qual o objetivo é procurar dissimular a origem ilícita de fundos obtidos através de atividade criminosa, distanciando a fonte do dinheiro da sua utilização, de forma que o dinheiro pareça ter uma origem lícita e que possa ser utilizado sem que haja suspeitas de sua origem (TIWARI; GEPP; KUMAR, 2020).

Para a lavagem de dinheiro, os criminosos geralmente convertem o dinheiro ilícito em criptomoedas e, em seguida, realizam várias transações entre diferentes carteiras de criptomoedas para dificultar o rastreamento. Eles também podem usar serviços de mistura de criptomoedas, conhecidos como *mixers* (MÖSER; BÖHME; BREUKER, 2013), que mesclam as transações de diferentes usuários para tornar mais difícil a identificação da origem do dinheiro.

Nesse contexto, alguns estudos já foram desenvolvidos, sendo que o presente trabalho tem como inspiração principal a pesquisa intitulada "*Detection of illicit accounts over the Ethereum blockchain*" de Farrugia, S., Ellul, J., & Azzopardi, G. (2020), que apresenta uma análise detalhada sobre a detecção de contas ilícitas na rede Ethereum. A citada pesquisa utilizou técnicas de aprendizado de máquina para identificar padrões de transações suspeitas e, assim, mapear contas que possivelmente estariam envolvidas em atividades ilegais. Os resultados obtidos na pesquisa mostraram-se bastante promissores, indicando que as técnicas utilizadas podem ser úteis para aprimorar a segurança e a transparência na rede Ethereum.

Assim, a motivação para a presente pesquisa é a crescente utilização de criptomoedas e blockchain no cenário global e a necessidade de aprimorar as técnicas de segurança nesse ambiente. O objetivo da pesquisa é investigar em que medida é possível identificar endereços de criptomoedas suspeitos de envolvimento com a prática da lavagem de dinheiro.

O tema abordado na pesquisa é novo e desafiador, frente à pouca regulamentação (TELLES, 2018) e pseudoanonimidade oferecida pela tecnologia, o que torna a identificação dessas contas um desafio. Ademais, a pesquisa envolve conceitos em discussão e uma abordagem tecnológica com aprendizado de máquina relativamente não usual na ciência social, quando comparado aos modelos matemáticos tradicionais de pesquisa. Além disso, a legislação recém-criada e ainda em construção para as criptomoedas e a falta de padronização nas

transações tornam o processo de detecção de contas ilícitas na rede Ethereum mais complexo. Com isso, tem-se por entendimento que a pesquisa agrega valor acadêmico e prático, contribuindo para o aprimoramento da segurança e transparência no uso de criptomoedas e das redes em blockchain, além de oferecer uma resposta às novas legislações que determinam o monitoramento e fiscalização dessas novas práticas no mercado financeiro.

Feito esse introito, a dissertação está organizada em cinco capítulos, dos quais o capítulo 1 é a presente introdução, no qual é feito uma breve apresentação das criptomoedas. O capítulo 2 é destinado ao referencial teórico da dissertação, no qual consta um breve aprofundamento sobre a blockchain Ethereum e reflexões em um contexto criminal envolvendo a lavagem de dinheiro. O capítulo 3 é dedicado ao método de pesquisa adotado inspirado no artigo acadêmico escrito por Farrugia, Ellul, Azzopardi (2020), contendo detalhes da coleta de dados e da análise dos dados. Já no capítulo 4 é feita apresentação dos resultados e discussão dos principais achados, junto da aplicação do modelo LightGBM, sendo que o fechamento da pesquisa é realizado no capítulo 5 com as conclusões.

2 REFERENCIAL TEÓRICO

Esta seção teve por objetivo oferecer uma revisão de literatura. Como ponto de partida, o leitor encontra uma breve contextualização sobre o conceito de blockchain e as criptomoedas mais populares e conhecidas, o Bitcoin e o Ethereum. Em seguida, discorreu-se sobre aspectos relacionados à Rede Ethereum e sua criptomoeda, o Ether, apresentando um panorama com mecanismos usualmente empregados por criminosos no que se refere ao branqueamento de capitais em ambiente blockchain. Encerra esta seção uma abordagem sobre estudos que se valeram da abordagem com máquina de aprendizagem para discutir esses fenômenos.

2.1 A tecnologia Blockchain

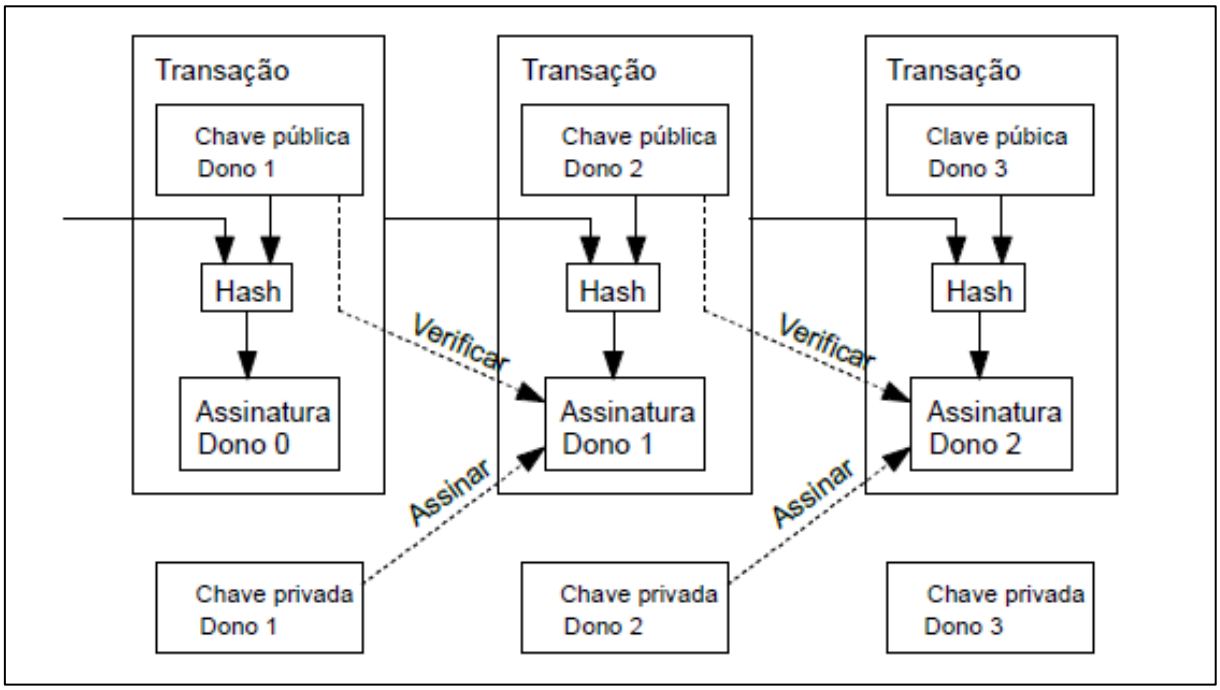
A ideia da tecnologia blockchain ficou mundialmente conhecida em 2008, com a publicação do artigo de título "*Bitcoin: A Peer-to-Peer Electronic Cash System*", trabalho que teve como autor Satoshi Nakamoto (pseudônimo). A ideia por trás dessa tecnologia envolvia a utilização de uma forma segura, descentralizada e transparente de registrar e verificar transações.

Blockchain, ou em tradução literal "corrente de blocos", traz a ideia de que novas informações ou transações são agrupadas em blocos e adicionados em uma cadeia de forma sequencial, cronológica e verificada por uma rede de computadores distribuídos, também chamados de nós da rede. Cada bloco contém um registro criptograficamente seguro das transações que ocorreram em um determinado período, e uma vez adicionado um novo bloco à cadeia de blocos, esse bloco não pode ser alterado sem a concordância de toda a rede, o que garante a segurança e a confiança nas transações registradas. Isso significa que o blockchain é um registro digital compartilhado e descentralizado, que é atualizado em tempo real e armazena informações de forma segura, sendo que sua rede é baseada em computadores que compartilham e validam transações por meio de consenso distribuído, o que inviabiliza fraudes ou meios de burlar a rede (NAKAMOTO, 2008), a Figura 1 ilustra esse esquema de funcionamento.

Uma forma mais simples de explicar o funcionamento desta tecnologia, é imaginar que a Blockchain é um grande livro de registro digital, onde as informações são registradas de forma segura e imutável. Esse livro de registro é compartilhado entre muitos computadores em diferentes lugares, formando uma rede descentralizada. A cada transação ou evento importante, uma nova página é adicionada ao livro de registro, e todas as cópias da rede são atualizadas simultaneamente, isso significa que todas as partes envolvidas têm acesso às mesmas informações e podem verificar a validade das transações.

O que torna a tecnologia blockchain especial é que as informações registradas são à prova de alterações, ou seja, uma vez que uma página é adicionada ao livro de registro, ela não pode ser apagada ou alterada facilmente, isso garante a segurança e a integridade dos dados e oferece uma forma confiável de armazenar e transferir informações, promovendo a transparência e a confiança entre as partes envolvidas.

Figura 1 - Funcionamento dos “blocos” dentro de uma blockchain



Fonte: Bitcoin: A Peer-to-Peer Electronic Cash System (NAKAMOTO, 2008)

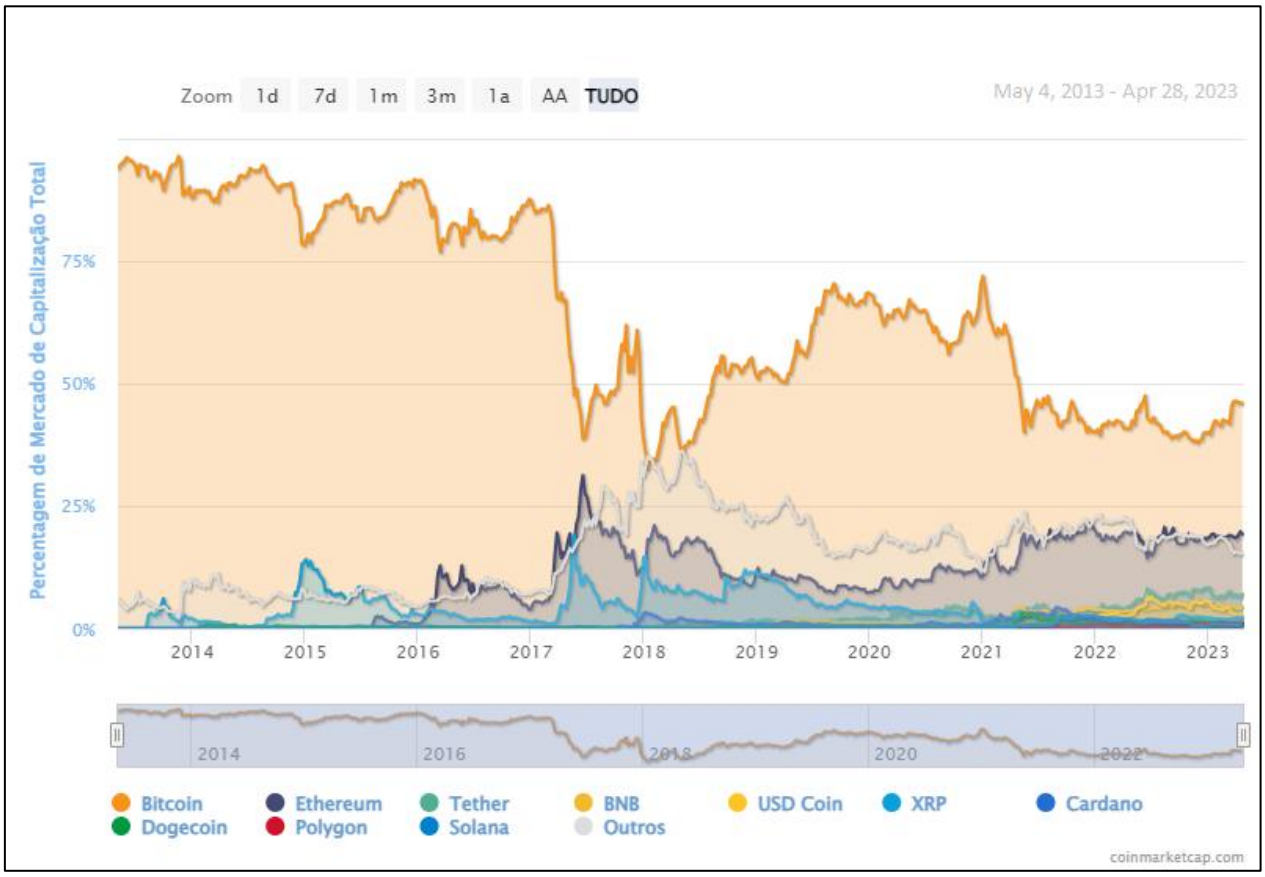
A tecnologia blockchain é conhecida por ser utilizada para as criptomoedas, como o Bitcoin, a mais conhecida delas. De acordo com Kadoo e Sodi (2023), o Bitcoin representa uma rede descentralizada que permite a transferência de dinheiro digital de forma segura e sem a necessidade de um intermediário confiável, possibilitando a circulação do dinheiro digital sem a intermediação de uma instituição financeira e governos, que cobram tarifas e impostos sobre a circulação financeira. Para os autores, o Bitcoin possui vantagens em relação aos sistemas de pagamentos existentes, incluindo a baixa taxa de transação, a rapidez e a privacidade, além de ser uma ferramenta a mais na proteção contra a inflação.

O mercado formado pelas criptomoedas apresenta uma capitalização de aproximadamente US$ 1,2 trilhão[1]. Assim, se o mercado de criptomoedas fosse um país, teria

[1] Informações retiradas do site https://www.coingecko.com/ (04/05/2023)

o 16º maior PIB (Produto Interno Bruto) no ranking do Banco Mundial (2021)[2], estando a frente de países como a Arábia Saudita e a Argentina. Dentre as criptomoedas, o Bitcoin ocupa o maior peso com 47,18% de participação em *market share*, enquanto a criptomoeda Ethereum, a segunda maior, corresponde a fatia de 19,28% do mercado cripto[3]. A Figura 2 ilustra a participação no mercado de criptomoedas.

Figura 2 - Principais Criptoativos por Porcentagem da Capitalização de Mercado Total



Fonte: coinmarketcap.com (data de referência: 04/05/2023).

Na visão de Pilkington (2015), a aplicação blockchain pode ir muito além do Bitcoin e das criptomoedas, como a aplicação da tecnologia em contratos inteligentes, provimento de identidades digitais baseados na tecnologia, sistemas de votação e o aumento da transparência na cadeia logística de produtos, dentre outros exemplos.

[2] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?most_recent_value_desc=true
[3] https://coinmarketcap.com (28/04/2023)

2.2 O Ethereum

Dentro do mundo das criptomoedas, Buterin (2014) defende que o Ethereum é uma plataforma de contrato inteligente (*smart contracts*) que permite a criação de contratos autônomos, programáveis e autoexecutáveis. Contratos inteligentes são programas de computador autônomos que são executados na blockchain da Ethereum, eles são projetados para automatizar e facilitar a execução de contratos digitais, eliminando intermediários e possibilitando transações seguras e transparentes. Na rede Ethereum esses contratos inteligentes, são armazenados e executados em cada nó da rede, eles contêm as regras e condições para a execução de um contrato, como as ações a serem tomadas quando determinadas condições forem cumpridas, apresentando inúmeras aplicações importantes como indicado no trabalho de Zou *et al.*, (2021).

O autor explica que os contratos inteligentes no Ethereum são capazes de automatizar processos, garantir a execução de acordos entre duas partes e, assim, reduzir a necessidade de intermediários em algumas transações. Suas aplicações iriam além das criptomoedas e poderiam ser usadas em várias áreas, como finanças, logística, votação, jogos online e identidade digital.

De acordo com o site especializado Coinmarketcap[4], o Ethereum, caracterizado pela sigla “ETH”, é a segunda maior criptomoeda em capitalização de mercado, com o valor aproximado de 229 bilhões de dólares americanos. Para os gestores da consultoria Elliptic[5], esse montante seria, em parte, decorrente da tentativa de alguns indivíduos no sentido de usar o Ethereum para lavar dinheiro movimentando fundos por meio de várias carteiras para esconder a origem dos recursos ou usar bolsas descentralizadas (*Decentralized Exchanges* - DEX), convertendo uma criptomoeda em outra para ofuscar ainda mais a fonte dos fundos. Além disso, os consultores entendem que contratos inteligentes baseados em Ethereum podem ser usados para facilitar transações financeiras complexas que podem ser difíceis de rastrear.

Cabe aclarar que endereços de Ethereum são uma combinação de letras e números gerados por algoritmos de criptografia, semelhantes aos endereços de e-mail, e permitem que pessoas enviem e recebam ETH e tokens. Os endereços são únicos e podem ser gerados por qualquer pessoa que tenha uma carteira Ethereum. Além disso, o Ethereum fornece uma plataforma poderosa para desenvolvedores criarem aplicativos descentralizados, possibilitando

[4] https://coinmarketcap.com, 2023
[5] https://www.elliptic.co/blog/money-laundering-through-dexs-and-mixers

uma nova geração de sistemas e organizações descentralizadas, com o potencial de expandir seu alcance e se tornar tão grande quanto o Bitcoin (BUTERIN, 2014).

2.3 Mundo cripto e a prática da lavagem de dinheiro

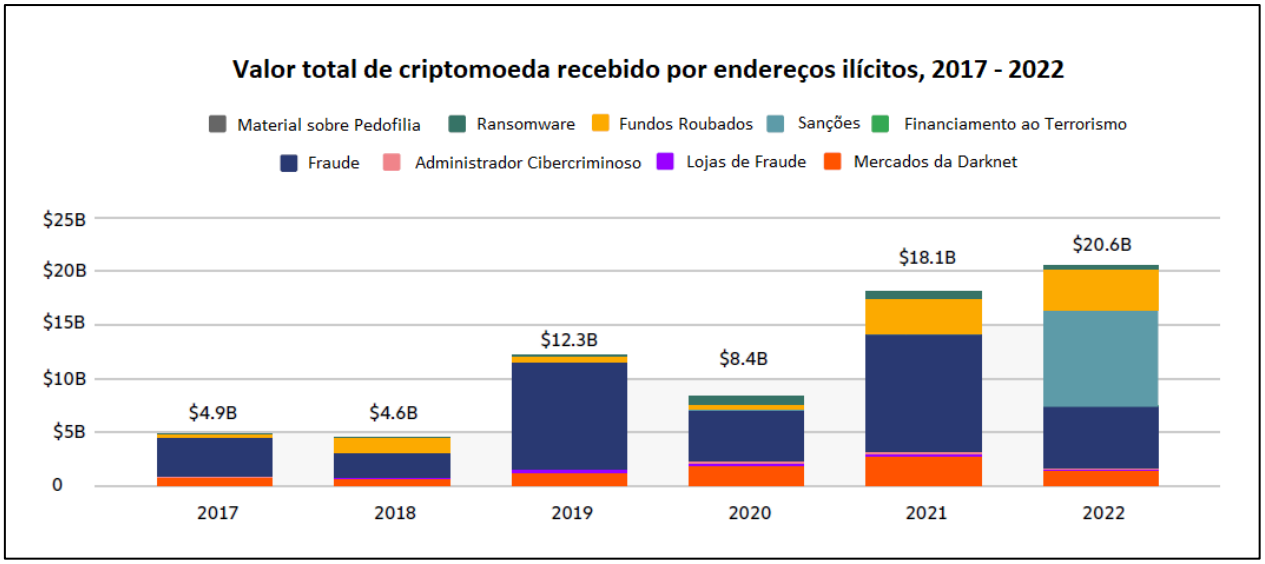
O mundo cripto tem sido amplamente debatido e discutido por sua inovação tecnológica. Novas formas de negociação financeira foram criadas, tornando-se cada vez mais presente esse mundo em nossas vidas. No entanto, assim como em qualquer outra área do mercado financeiro, a prática de crimes também pode ocorrer no âmbito das criptomoedas, incluindo lavagem de dinheiro, fraudes, evasão fiscal, corrupção e financiamento ao terrorismo (AKARTUNA; JOHNSON; THORNTON, 2022).

Nos primórdios do surgimento e utilização do Bitcoin houve muita especulação sobre a relação entre o seu uso e atividades ilícitas, como o terrorismo, pornografia infantil e a compra e venda de drogas. Nica *et al.* (2017) afirmam que o caso *Silk Road* foi o mais famoso escândalo, consistindo em um mercado negro online de drogas e substâncias farmacêuticas que não apenas aceitava, mas estabelecia como meio de pagamento das transações o Bitcoin.

A dimensão e a utilização dos criptoativos levou ao maior interesse por parte de criminosos que viam nesse mercado uma ferramenta potencial para facilitar a prática de crimes financeiros (TELLES, 2018), como a lavagem de dinheiro, golpes e fraudes que envolviam as transações por meio de criptomoeda, principalmente pelo fato de as transações serem pseudoanônimas.

De acordo com pesquisa da Chainalysis Inc. (2023), uma das maiores empresa de análise de blockchain, os crimes financeiros envolvendo criptomoedas atingiram a cifra de US$ 20,6 bilhões em 2022, em franco crescimento nos últimos anos, tendo entre os principais tipos de crime fraudes diversas, como aqueles relacionadas à lavagem de dinheiro e corrupção, além de roubo de criptoativos, sanções e crimes cometidos na *darknet*.

Figura 3 - Relação de valores recebidos por endereços ilícitos



Fonte: The 2023 Crypto Crime Report CHAINALYSIS INC. (2023)

Para Connolly e Wall (2019), as criptomoedas podem ser utilizadas como forma de pagamento de ataques *ransomware*, que é um software malicioso que criptografa os arquivos do computador da vítima e impede o acesso a eles, exigindo um pagamento (resgate) em troca da chave de descriptografia, sendo o objetivo dos criminosos extorquir dinheiro das vítimas, muitas vezes ameaçando destruir ou publicar informações sensíveis ou importantes.

Já em mercados clandestinos como a *darknet*, grandes quantidades de bens e serviços – como drogas, armas e ataques DDoS – são comprados e vendidos usando criptomoeda como método de pagamento a fim de ocultar a identificação das partes envolvidas. Em mercados clandestinos online, criptomoedas são, portanto, consideradas a forma de pagamento preferida dos criminosos. (KRUISBERGEN *et al.*, 2019)

Para utilizar o dinheiro obtido a partir dos crimes mencionados anteriormente, é necessário “limpar” o dinheiro para ocultar sua origem ilícita. O uso de criptoativos para a lavagem de dinheiro é uma prática que tem crescido nos últimos anos por conta da natureza descentralizada e o pseudoanonimato das moedas digitais (VAN WEGBERG; OERLEMANS; VAN DEVENTER, 2018). Os criminosos podem usar criptomoedas para transferir fundos ilícitos de um lugar para outro de forma relativamente anônima e sem deixar muitos rastros. Além disso, as transações em criptomoedas não possuem limites, o que permite a movimentação de grandes quantias de dinheiro sem levantar suspeitas e sem a necessidade de intermediários, como banco e governos, não tendo limitações transfronteiriças (MÖSER; BÖHME; BREUKER, 2013).

No que se refere à lavagem de dinheiro, tipificada no ordenamento nacional com a Lei 9.613/1998, existem diversos tipos e técnicas relacionadas à essa prática ilegal, as quais podem ser didaticamente descritas em três etapas principais, conhecidas como "colocação" (*placement*), "ocultação" (*layering*) e a "integração" (*integration*) (FATF, 2005). Essas etapas representariam o processo pelo qual os valores provenientes de atividades ilícitas são dissimulados e integrados ao sistema financeiro tradicional com a aparência de terem sido obtidas de forma legal.

No entendimento de Fanusie e Robinson (2018), dentre os crimes financeiros, o uso de criptomoedas pode contribuir em diversas etapas do processo de lavagem de dinheiro, uma vez que sua natureza descentralizada e sua característica de permitir transações semianônimas são potenciais interesses de delinquentes voltados à lavagem de dinheiro, em particular, nas fases de ocultação e integração.

Segundo Irwin *et al.* (2011), lavadores de dinheiro têm preferência por técnicas desprendidas, como o uso de *smurfing,* procedimento que significa dividir altos valores ilícitos em quantias menores para contornar os limites regulatórios dos órgãos de controle e fiscalização, reduzindo o nível de desconfiança do sistema sobre uma transação (EIFREM, 2019).

Existem, ainda, outras técnicas como o fracionamento de recursos, ou seja, a quantidade transferida frequentemente é pequena e são feitas inúmeras transferências, para evitar o chamamento de atenção, podendo utilizar contas de passagem (*pooling accounts*) para camuflar a origem dos valores, conforme explica Lima (2021).

No entanto, é importante ressaltar que a utilização de criptomoedas não é uma garantia de sucesso na lavagem de dinheiro, pois as transações podem ser rastreadas através da análise da blockchain e outras técnicas de investigação. Ademais, as autoridades estão cada vez mais conscientes dos riscos associados às criptomoedas e estão implementando medidas regulatórias e de combate à lavagem de dinheiro para prevenir o uso ilícito desses ativos (CAMPBELL-VERDUYN, 2018). Nesse sentido, Telles (2018) pondera que a regulamentação das criptomoedas ainda é incipiente e as autoridades têm um desafio cada vez maior em identificar e prevenir essas atividades ilícitas. No cenário atual, já existem iniciativas como a de Aziz *et al.*, (2022), nas quais são utilizadas ferramentas de aprendizado de máquina, na detecção de fraudes envolvendo blockchain e criptomoedas.

Para enfrentar esse desafio, recentes inovações legislativas, como a Lei nº 14.478/2022, tentam disciplinar a prestação de serviços de ativos virtuais e regulamentar as prestadoras de serviços de ativos virtuais, o que inclui a compreensão e regulação desses novos ativos no que

tange à prática de crimes, como o seu uso para a lavagem de dinheiro, conforme disposto no art. 4º, inciso VII, na referida lei:

> Art. 4º A prestação de serviço de ativos virtuais deve observar as seguintes diretrizes, segundo parâmetros a serem estabelecidos pelo órgão ou pela entidade da Administração Pública federal definido em ato do Poder Executivo:
> I - livre iniciativa e livre concorrência;
> II - boas práticas de governança, transparência nas operações e abordagem baseada em riscos;
> III - segurança da informação e proteção de dados pessoais;
> IV - proteção e defesa de consumidores e usuários;
> V - proteção à poupança popular;
> VI - solidez e eficiência das operações; e
> VII - **prevenção à lavagem de dinheiro e ao financiamento do terrorismo e da proliferação de armas de destruição em massa, em alinhamento com os padrões internacionais**. (g.n.)

Outrossim, a Lei nº 14.478/2022 também alterou a Lei 9.613/1998, quando promoveu menção a utilização de ativo virtual como agravante do crime de lavagem de ativos:

> Art. 12. A Lei nº 9.613, de 3 de março de 1998, passa a vigorar com as seguintes alterações:
> "§ 4º A pena será aumentada de 1/3 (um terço) a 2/3 (dois terços) se os crimes definidos nesta Lei forem cometidos de forma reiterada, por intermédio de organização criminosa ou **por meio da utilização de ativo virtual**." (g.n.)

Ainda, motiva mencionar o Decreto nº 11.563/2023, o qual regulamentou a Lei nº 14.478/2022 e estabeleceu o Banco Central do Brasil como órgão da administração pública federal responsável por disciplinar o funcionamento das prestadoras de serviços de ativos virtuais, ficando inclusive responsável pela supervisão das referidas prestadoras.

2.4 Modelos de aprendizado de máquina na identificação de fraudes

A aplicação de técnicas de *machine learning* na pesquisa com o enfoque na detecção de lavagem de dinheiro é um campo de estudo em crescimento, como na pesquisa de Jullum *et al.* (2020). Os autores propuseram um modelo aplicado em uma base de dados de transações financeiras do maior banco da Noruega em busca de sinais desse crime. De acordo com Ruiz e Angelis (2022), a utilização desses métodos no ambiente das criptomoedas tem sido amplamente explorada nos últimos anos, pois o aprendizado de máquina oferece uma abordagem promissora para analisar e extrair percepções valiosas dos dados relacionados às transações em criptomoedas, incluindo a detecção de atividades suspeitas, como a lavagem de dinheiro.

O aprendizado de máquina (ou "*machine learning*" em inglês) é uma subárea da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem que computadores “aprendam” a partir de dados, com a combinação de três fatores: a representação, a avaliação e a otimização (DOMINGOS, 2012). Em vez de programar explicitamente uma máquina para executar uma tarefa específica, o aprendizado de máquina usa algoritmos para aprender a partir de exemplos e dados históricos, e com isso fazer previsões, tomar decisões com base nesse aprendizado e fornecer percepções valiosas sobre os dados.

Lorenz *et al.* (2020) e Alarab, Prakoonwit e Nacer (2020) pesquisaram modelos com algoritmos de *machine learning* para detectar lavagem de dinheiro na blockchain do Bitcoin. Os resultados obtidos demonstraram a eficácia dessa abordagem no combate a esse tipo de crime envolvendo criptoativos. Além disso, Ruiz e Angelis (2022) destacam a relevância do uso de técnicas de aprendizado de máquina nesse contexto, ressaltando seus benefícios no combate à lavagem de dinheiro na blockchain.

Especialmente no tocante à rede Ethereum, Ibrahim, Elian e Ababneh (2021) investigaram o uso de *machine learning* na detecção de carteiras ilícitas com os modelos *decision tree (j48), Random Forest* e *K-nearest neighbors*, tendo por objetivo automatizar a seleção de carteiras suspeitas. A utilização de modelos de aprendizado de máquina está sendo usada cada vez mais na detecção de fraudes, sendo uma importante ferramenta devido ao grande volume de informações oriundas das transações do mundo cripto, podemos citar também o trabalho de Aziz *et al.*, (2023), que se utilizada dessas ferramentas na detecção de transações fraudulentas em contratos inteligentes na rede Ethereum, e o trabalho da mesma autora (AZIZ *et al.*, 2022), no qual foi utilizado uma abordagem com aprendizado de máquina para detectar fraudes na rede Ethereum.

Por sua vez, o estudo conduzido por Farrugia, Ellul e Azzopardi (2020), que também abordou a detecção de contas ilícitas na blockchain Ethereum, utilizou o modelo XGBoost sobre uma base de dados composta pela combinação de duas fontes, o Etherscamdb e um cliente Geth local conectado à rede Ethereum para contas de golpes e contas normais, respectivamente. Essa abordagem teve como objetivo destacar as carteiras relacionadas à blockchain Ethereum que estão mais indicadas a envolvimento com transações ilícitas.

3 MÉTODO DE PESQUISA

O objetivo da presente pesquisa é aprofundar a investigação sobre a viabilidade de identificação de endereços de criptomoedas suspeitos de envolvimento com a prática da lavagem de dinheiro com auxílio de ferramentas de *machine learning*, com foco no blockchain Ethereum. O trabalho tem como inspiração o artigo acadêmico escrito por Steve Farrugia, Joshua Ellul e George Azzopardi (2020), intitulado *Detection of Illicit Accounts over the Ethereum Blockchain*, no qual foi utilizado o modelo de aprendizagem de máquina conhecido por XGBoost (Extreme Gradient Boosting) para a detecção de possíveis padrões em endereços na rede Ethereum envolvidos com atividade ilícitas.

Por oportuno, esclarece-se que o XGBoost é um modelo de classificação que usa a técnica de *boosting* para combinar vários modelos mais simples em um modelo mais forte e preciso. O XGBoost usa árvores de decisão como os "*weak learners*" e, através do processo de *boosting*, ajusta cada árvore para corrigir os erros dos modelos anteriores. Isso resulta em um modelo mais preciso e robusto, capaz de lidar com soluções de problemas de escala global usando poucos recursos (CHEN; GUESTRIN, 2016). Em resumo, o XGBoost é um modelo de aprendizado de máquina que usa uma técnica de combinação de árvores de decisão para criar um modelo de classificação preciso e escalável.

3.1 Coleta de dados

A base de dados usada neste trabalho foi obtida a partir do estudo de Farrugia et al. (2020), dataset disponibilizado de forma pública em link[6] da plataforma web Github (site utilizado mundialmente para hospedagem de código-fonte e projetos compartilhados). Farrugia et al. (2020) esclarecem que o dataset da pesquisa foi obtido pela combinação de duas fontes: Etherscamdb para contas vinculadas a golpes; e um cliente Geth local conectado à rede Ethereum para contas sem notações de fraudes.

Nesse ínterim, importa elucidar que o Etherscamdb é um banco de dados que possui endereços de carteiras de criptomoedas que estão associados a atividades fraudulentas na rede Ethereum. O banco de dados é mantido por uma comunidade de voluntários que monitoram a rede Ethereum em busca de golpes e fraudes, e adicionam os endereços identificados ao banco de dados (FARRUGIA et al., 2020; CHAVES, 2021; RODRIGUES *et al.*, 2021). Já o cliente Geth é um software livre e de código aberto que pode ser baixado diretamente do site oficial da

[6] https://github.com/sfarrugia15/Ethereum_Fraud_Detection/blob/master/Account_Stats/Complete.csv

Ethereum Foundation. O código-fonte do cliente Geth está disponível no Github[7] e pode ser utilizado por qualquer pessoa ou organização interessada em se conectar à rede Ethereum e interagir com ela (FARRUGIA; ELLUL; AZZOPARDI, 2020).

Registra-se que o acesso ao banco de dados se deu em 13 de abril de 2023, sendo que o dataset original foi criado em 17 de abril de 2019. O dataset possui 2179 contas apontadas como ilícitas pela comunidade Ethereum e 2502 contas não apontadas como ilícitas, totalizando 4681 endereços na rede Ethereum.

3.2 Análise de dados

Avançando a partir do estudo de Steven Farrugia, Joshua Ellul e George Azopardi (2020), a presente pesquisa se propõe a utilizar o modelo conhecido como LightGBM (*Light Gradient Boosting Machine*). KE *et al.* (2017) explicam que este modelo é um algoritmo de aprendizado de máquina que também se baseia na técnica de *boosting* para criar modelos de classificação e regressão, mas com uma abordagem que busca a otimização do processo de aprendizado, chamada de "*histogram-based*", cuja finalidade é reduzir a complexidade computacional e melhorar a eficiência do treinamento. O LightGBM seria capaz de lidar com conjuntos de dados grandes e complexos, usando técnica de amostragem por nó para dividir os dados em diferentes camadas de folhas. Ademais, esse modelo tem sido usado com sucesso em vários problemas de aprendizado de máquina, incluindo classificação de imagem (MICHAEL et al., 2022), previsão de preços de imóveis (M. JOHN et al., 2022) e detecção de fraudes em cartões de crédito (HUANG, 2020).

Para avaliação do modelo LightGBM, serão utilizados testes de diagnóstico semelhantes aos utilizados no estudo base, conforme elencado e descrito a seguir de acordo com as definições do livro de Patterson e Gibson (2017):

(i) Acurácia: medida da proporção de previsões corretas feitas pelo modelo em relação ao total de previsões realizadas, ou seja, a taxa de sucesso do modelo em classificar corretamente os exemplos;

$$\text{Acurácia} = (TP + TN) / (TP + FP + FN + TN)$$

[7] https://github.com/ethereum/go-ethereum

(ii) Precisão: é uma medida de desempenho que indica, dentre todas as classificações de classe Positivo que o a proporção de identificações positivas que realmente estavam corretas;

$$\text{Precisão} = TP / (TP + FP)$$

(iii) Revocação ou sensibilidade: também conhecida como recall ou taxa de verdadeiros positivos, é uma métrica de desempenho usada para avaliar a taxa de verdadeiros positivos em relação ao total de amostras positivas reais;

$$\text{Revocação} = TP / (TP + FN)$$

(iv) F1-score: é uma métrica de desempenho que combina a precisão e a revocação de um modelo de classificação em uma única medida; e

$$F1 = 2TP / (2TP + FP + FN)$$

(v) ROC-AUC (*Receiver Operating Characteristic - Area Under the Curve*): é uma métrica comumente usada para avaliar a capacidade de um modelo de classificação em distinguir entre classes positivas e negativas, ou seja, é uma medida utilizada para constatar a qualidade do desempenho de um modelo de classificação, é calculado traçando a curva ROC, que representa a taxa de verdadeiros positivos (TPR) em relação à taxa de falsos positivos (FPR) para diferentes limiares de classificação. A área sob essa curva (AUC) é então calculada para obter o valor do ROC-AUC.

Além dos cinco testes de diagnóstico, será elaborada uma matriz de confusão, que é a representação tabular usada para visualizar o desempenho de um modelo de aprendizado de máquina em problemas de classificação.

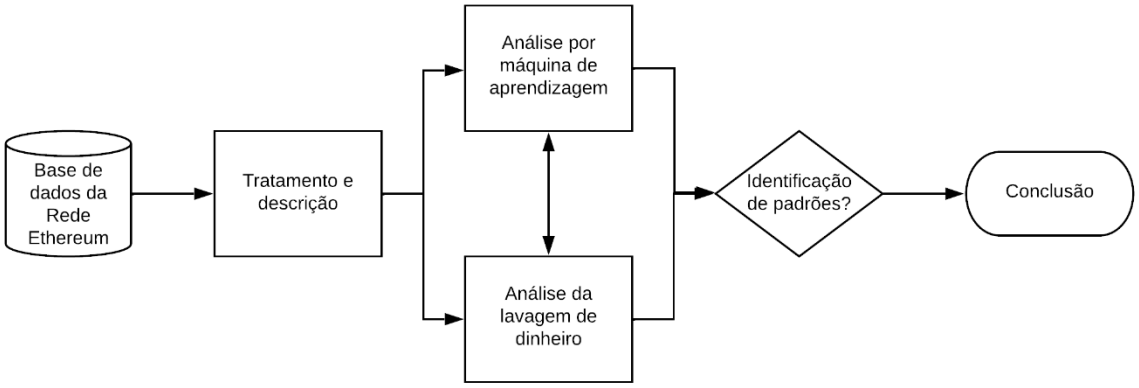
Figura 4 – Matriz de confusão

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo ou *True Positive* (TP)	Falso Negativo ou *False Negative* (FN)
	Não	Falso Positivo ou *False Positive* (FP)	Verdadeiro Negativo ou *True Negative* (TN)

Fonte: Adaptado de Patterson e Gibson (2017).

Após avaliação do modelo do ponto de vista da qualidade do algoritmo de máquina de aprendizagem, desta vez inspirado em Lima (2021), será realizada análise no sentido de verificar operações e situações que habitualmente são taxadas como indícios de lavagem de dinheiro. Assim, será considerada a possiblidade de reconhecer as fases do processo de lavagem de dinheiro nas contas analisadas pelo modelo. A análise dos dados seguiu o fluxograma demonstrado na Figura 5.

Figura 5 - Fluxo da análise de dados



Fonte: Adaptado de Lima (2021).

4 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados das análises realizadas, iniciando pela análise descritiva dos dados, seguido dos procedimentos de preparação dos dados (pré-processamento), depois pela análise dos resultados do modelo de máquina de aprendizagem e da discussão dos achados sob o prisma da lavagem de dinheiro, especialmente no que tange aos comportamentos que sugiram a prática criminosa.

4.1 Análise descritiva dos dados

O banco de dados original possuía 49 atributos. O atributo "*Address*" se refere ao identificador da carteira, o endereço na rede Ethereum, portanto não foi usado como entrada do classificador, e "*FLAG*" que é o atributo-alvo (endereço lícito ou ilícito), elemento necessário para treinamento do algoritmo classificador e avaliação do desempenho do modelo.

Na descrição contida no artigo, Farrugia, Ellul e Azzopardi (2020) apontam que usaram 42 atributos. No entanto, no arquivo disponibilizado por eles, foram encontrados 47 atributos[8], sendo 5 elementos[9] encontrados a mais. Por não estarem indicados e documentados na descrição da base original, considerando a incerteza da natureza dos referidos atributos e a falta de evidência de que tais dados seriam, foi decidido "por prudência" pela não utilização desses 5 atributos na análise em questão.

A Tabela 1 é composta dos 42 atributos, além do atributo alvo e o endereço da carteira[10] para demonstração do conjunto completo a ser analisado.

Tabela 1 – Conjunto completo de atributos que serão utilizados

	Conjunto Completo de Atributos que Serão Utilizados	
	Atributos Extraídos	**Descrição**
1	**Address**	Endereço da carteira
2	**FLAG**	Atributo alvo (endereço lícito ou ilícito)
3	**Avg_min_between_sent_tnx**	Tempo médio entre as transações enviadas para a conta em minutos
4	**Avg_min_between_received_tnx**	Tempo médio entre as transações recebidas por conta em minutos
5	**Time_Diff_between_first_and_last(Mins)**	Diferença de tempo entre a primeira e a última transação (Mins)

[8] Não foram considerados o atributo "*Address*" e "*FLAG*".
[9] "ERC20_uniq_sent_addr.1", "ERC20_avg_time_between_rec_2_tnx", "ERC20_avg_val_sent_contract", "ERC20_max_val_sent_contract" e "ERC20_min_val_sent_contract".
[10] "*Address*" e "*FLAG*".

	Conjunto Completo de Atributos que Serão Utilizados	
	Atributos Extraídos	**Descrição**
6	**Sent_tnx**	Número total de transações normais enviadas
7	**Received_tnx**	Número total de transações normais recebidas
8	**Number_of_Created_Contracts**	Número total de transações de contratos criados
9	**Unique_Received_From_Addresses**	Total de endereços únicos dos quais a conta recebeu transações
10	**Unique_Sent_To_Addresses**	Total de endereços únicos dos quais a conta enviou transações
11	**Min_Value_Received**	Valor mínimo em Ether já recebido
12	**Max_Value_Received**	Valor máximo em Ether já recebido
13	**Avg_Value_Received**	Valor médio em Ether já recebido
14	**Min_Val_Sent**	Valor mínimo de Ether já enviado
15	**Max_Val_Sent**	Valor máximo de Ether já enviado
16	**Avg_Val_Sent**	Valor médio de Ether já enviado
17	**Min_Value_Sent_To_Contract**	Valor mínimo de Ether enviado para um contrato
18	**Max_Value_Sent_To_Contract**	Valor máximo de Ether enviado para um contrato
19	**Avg_Value_Sent_To_Contract**	Valor médio de Ether enviado para contratos
20	**Total_Transactions(Including_Tnx_to_Create_Contract)**	Número total de transações
21	**Total_Ether_Sent**	Total de Ether enviado para o endereço da conta
22	**Total_Ether_Received**	Total de Ether recebido para o endereço da conta
23	**Total_Ether_Sent_Contracts**	Total de Ether enviado para endereços de contrato
24	**Total_Ether_Balance**	Saldo total de Ether após transações aprovadas
25	**Total_ERC20_Tnxs**	Número total de transações de transferência de token ERC20
26	**ERC20_Total_Ether_Received**	Total de token ERC20 recebidos transações em Ether
27	**ERC20_Total_Ether_Sent**	Total de token ERC20 enviados transações em Ether
28	**ERC20_Total_Ether_Sent_Contract**	Total de transferências de token ERC20 para outros contratos em Ether
29	**ERC20_Uniq_Sent_Addr**	Número de transações de token ERC20 enviadas para endereços de contas únicas
30	**ERC20_Uniq_Rec_Addr**	Número de transações de token ERC20 recebidas de endereços únicos

	Conjunto Completo de Atributos que Serão Utilizados	
	Atributos Extraídos	**Descrição**
31	**ERC20_Uniq_Rec_Contract_Addr**	Número de transações de token ERC20 recebidas de endereços contratos únicos
32	**ERC20_Avg_Time_Between_Sent_Tnx**	Tempo médio entre transações enviadas de token ERC20 em minutos
33	**ERC20_Avg_Time_Between_Rec_Tnx**	Tempo médio entre as transações recebidas de token ERC20 em minutos
34	**ERC20_Avg_Time_Between_Contract_Tnx**	Tempo médio token ERC20 entre transações de token enviadas
35	**ERC20_Min_Val_Rec**	Valor mínimo em Ether recebido de transações de token ERC20 para conta
36	**ERC20_Max_Val_Rec**	Valor máximo em Ether recebido de transações de token ERC20 para conta
37	**ERC20_Avg_Val_Rec**	Valor médio em Ether recebido de transações de token ERC20 para conta
38	**ERC20_Min_Val_Sent**	Valor mínimo em Ether enviado de transações de token ERC20 para conta
39	**ERC20_Max_Val_Sent**	Valor máximo em Ether enviado de transações de token ERC20 para conta
40	**ERC20_Avg_Val_Sent**	Valor médio em Ether enviado de transações de token ERC20 para conta
41	**ERC20_Uniq_Sent_Token_Name**	Número de tokens ERC20 únicos transferidos
42	**ERC20_Uniq_Rec_Token_Name**	Número de tokens ERC20 únicos recebidos
43	**ERC20_Most_Sent_Token_Type**	Token mais enviado para conta via transação ERC20
44	**ERC20_Most_Rec_Token_Type**	Token mais recebido para conta via transação ERC20

Nota: Os atributos 3 a 42 são numéricos. Os atributos 43 e 44 são categóricos.

Foram observados cinco endereços duplicados, sendo que a conta 0xd624d046edbdef805c5e4140dce5fb5ec1b39a3c possuía a indicação (atributo "FLAG") de ilícita em uma de suas cópias e de lícita na outra. Os endereços duplicados foram excluídos da base de dados[11], sendo que a conta com sinalização de lícita e ilícita foi mantida apenas no grupo das ilícitas[12].

[11] Os endereços excluídos foram:
0xd624d046edbdef805c5e4140dce5fb5ec1b39a3c
0x75e7f640bf6968b6f32c47a3cd82c3c2c9dcae68
0x8271b2e8cbe29396e9563229030c89679b9470db
0x91337a300e0361bddb2e377dd4e88ccb7796663d
0x96fc4553a00c117c5b0bed950dd625d1c16dc894

[12] Essa opção foi um tanto arbitrária. A escolha decorre do entendimento de que se a conta foi incluída no Etherscamdb por algum membro da comunidade, provavelmente há indícios que ela possui relação com atividade ilícita, mesmo estando no rol de contas lícitas.

Além dos endereços duplicados citados acima, foram encontradas mais duas situações (conjuntos de atributos) que podem ajudar a reduzir a quantidade de atributos no modelo, tornando-o mais simples, interpretável e eficiente.

O primeiro conjunto, por se tratar de 4 atributos, sendo que cada um desses atributos possui o mesmo valor para todas as carteiras da amostra, não agrega diferença alguma ou informação adicional entre elas para a classificação que será feita pelo modelo. Assim, optou-se pela exclusão desses atributos na análise, conforme Tabela 2.

Tabela 2 – Atributos excluídos com valores constantes

Atributos com valores constantes para todas as carteiras
min_value_sent_to_contract,
max_val_sent_to_contract,
avg_value_sent_to_contract,
total_ether_sent_contracts

No segundo conjunto, o valor dos atributos para cada uma das carteiras era igual entre si, portanto, foi feita a utilização de somente um atributo desse segundo grupo e feita a remoção dos demais, conforme indicado na Tabela 3, a seguir:

Tabela 3 - Conjunto de atributos iguais entre si

Atributo a ser mantido	**Atributos com valores iguais entre si**
ERC20_avg_time_between_contract_tnx	ERC20_avg_time_between_sent_tnx ERC20_avg_time_between_rec_tnx ERC20_avg_time_between_contract_tnx

Nessa etapa, o dataset apresenta uma proporção de 46,5% de contas ilícitas, o que permite inferir ser uma base balanceada e sem a necessidade de usar técnicas de balanceamento em relação ao atributo-alvo.

4.2 Pré-processamento dos dados

A fase de pré-processamento tem o objetivo de realizar as transformações necessárias no conjunto de dados que possibilite seu melhor uso pelo modelo de *machine learning* proposto. Foram feitas substituições de valores nulos pela mediana no caso de atributos numéricos, para a preservação da distribuição dos dados e robustecer contra *outliers* extremos (LITTLE; RUBIN, 2020) e a transformação de dados categóricos em números. Dessa forma, o pré-processamento foi realizado considerando apenas o domínio dos atributos do conjunto de

treinamento, ao passo que no conjunto de teste foi realizada a simples reprodução dos passos adotados no conjunto de treinamento.

O processo de separação dos grupos de treinamento e de teste para posterior tratamento é essencial nos casos de utilização de *machine learning*, para segregar totalmente os grupos, evitando assim o *data leakage*, que consiste no erro no qual informações sobre a variável alvo vazam para o *input* do modelo durante o treinamento do modelo, informações que não estarão disponíveis em dados futuros.

Ao final foi elaborada a Tabela 4, indicando que restaram 38 atributos e 4676 contas que serviram como base para este trabalho.

Tabela 4 - Primeira etapa do pré-processamento

Etapa	Atributos	Contas
Base original disponibilizada pelo autor	49	4681
Remoção de divergências (atributos não indicados na descrição da base original)	-5	-
Remoção de divergências (contas duplicadas)	-	-5
Atributos com valores constantes (Tabela 2)	-4	-
Atributos iguais (Tabela 3)	-2	-
Base deste trabalho	38	4676

Ademais, foi necessário separar o conjunto de dados em duas partes, sendo a primeira parcela para o treinamento do modelo e o papel da segunda parcela, chamada de conjunto de teste, é avaliar o comportamento do modelo com dados obtidos do processo que se quer avaliar, mas que o modelo não teve acesso. A separação foi realizada usando-se uma amostragem aleatória dos conjuntos de treino e teste (TREVOR; ROBERT; FRIEDMAN, 2013).

4.2.1 Atributos numéricos

A cardinalidade refere-se ao número de elementos únicos em um conjunto de dados ou em um mesmo atributo (categoria). É uma medida da diversidade ou variedade dos valores presentes em um conjunto de dados, é frequentemente usada para avaliar a complexidade de um conjunto de dados, pois afeta diretamente a quantidade de informações únicas que podem ser extraídas dele. Diz-se que um atributo possui alta cardinalidade quando ele possui um número elevado de opções possíveis.

A proporção de nulos entre os atributos numéricos é 0% em 18 dos atributos e 17,7% nos outros 16. Além disso, a ocorrência de nulos é concomitante ao longo dos mesmos 16

atributos, ou seja, se em um desses 16 atributos é nulo, os outros 15 atributos também serão nulos.

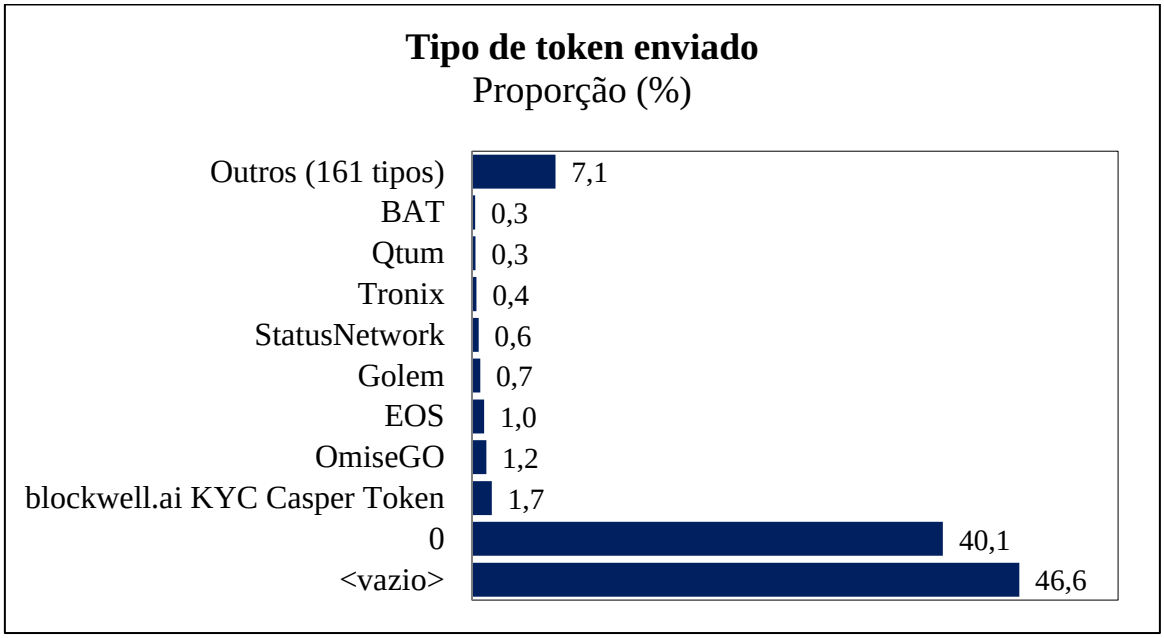
A estratégia adotada para os registros com valores nulos foi a de substituí-los por sua mediana antes de serem usados para treinar o modelo, a fim de não perdermos 17,7% da base com a deleção desses registros, pois ainda podemos contar com os outros 18 atributos numéricos não nulos.

Com isso, o único pré-processamento realizado nos atributos numéricos foi a substituição dos valores nulos pela mediana dos atributos do conjunto de treinamento. Para o conjunto de teste, ocorre a substituição dos valores nulos pela mediana obtida dos atributos do conjunto de treinamento.

4.2.2 Atributos categóricos

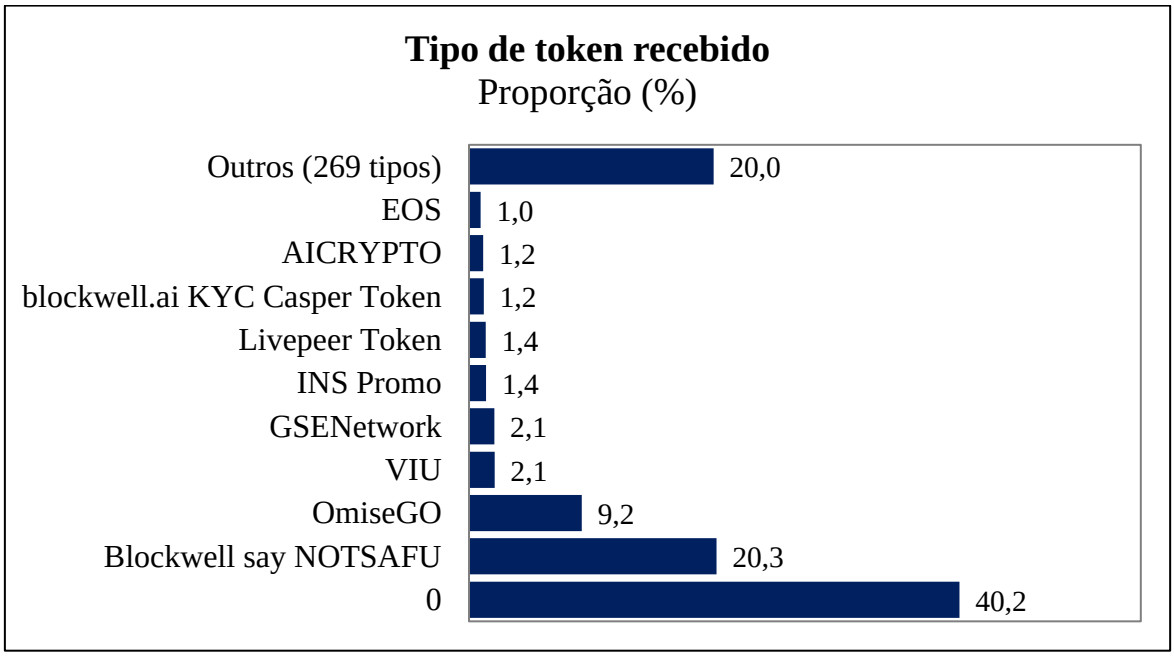
Há duas colunas (atributos) com valores categóricos: ERC20_most_rec_token_type e ERC20_most_sent_token_type. Os gráficos representados pela Figura 6 e Figura 7 mostram como a ocorrência dos tipos de token recebidos ou enviados é desbalanceada no conjunto de dados. Na Figura 6, as categorias representadas por “0” e “<vazio>”, categorias que indicam que na base não havia informação do tipo de token mais enviado naquela conta, representam 86,7% dos valores, enquanto as outras 169 diferentes categorias representam os 13,3% restantes. Já na Figura 7, ainda temos a hegemonia do valor “0”, porém, com 40,2% do total de ocorrências. Além disso, temos um cenário de tipos de tokens menos pulverizado com cerca de 278 tipos de tokens.

Figura 6 - Proporção de ocorrências dos valores mais frequentes para o atributo ERC20_most_sent_token_type.



Fonte: Autor.

Figura 7 - Proporção de ocorrências dos valores mais frequentes para o atributo ERC20_most_rec_token_type.



Fonte: Autor.

Tendo em vista o desbalanceamento das ocorrências dos tipos de token e a alta cardinalidade desses dois atributos categóricos, uma estratégia que pode ser tomada para introduzir esses atributos em um modelo é agrupar os tipos de tokens em grupos e trocar seu

respectivo grupo resultante em sua frequência de ocorrência, transformando-os em atributos numéricos.

No caso dos dois atributos categóricos presentes (“ERC20_most_rec_token_type” e “ERC20_most_sent_token_type”), é importante pontuar que os valores nulos, “0” ou “ “ (espaços em branco) foram substituídos pelo texto “*no information*”. Posteriormente, a estratégia adotada foi reduzir a cardinalidade através do agrupamento das categorias com ocorrência mais rara em um mesmo grupo chamado “raro”. Após o agrupamento, substituímos os respectivos valores categóricos pela frequência de ocorrência de cada uma das categorias ou classes.

Tabela 5 - Proporção das categorias em ERC20_most_rec_token_type após o agrupamento das categorias com ocorrência rara.

ERC20_most_rec_token_type	Proporção (%)	Acumulado (%)	Substituído por
no information	51,4	51,4	0,51421384
Blockwell say NOTSAFU	16,7	68,1	0,16704403
Raro	31,9	100,0	0,31874214

Tabela 6 - Proporção das categorias em ERC20_most_sent_token_type após o agrupamento das categorias com ocorrência rara.

ERC20_most_sent_token_type	Proporção (%)	Acumulado (%)	Substituído por
no information	89,0	89,0	0,89006289
Raro	11,0	100,0	0,10993711

4.3 Aplicação do modelo

O modelo foi treinado com 3974 registros (85% do total da base) e validado com 702 registros (15% do total da base de 4676 registros), após a exclusão dos registros com divergências, conforme relatado em 4.1. Essa divisão foi adotada com base no trabalho de Joseph (2022), que em seu exemplo utilizou a mesma proporção considerando uma base de dados com características semelhantes a deste trabalho e dados trazidos por Larsen e Goutte (1999), sendo indicado que a razão deve ficar mais próximo de 100%, quanto maior for o número de dados. Cabe ressaltar que antes do treinamento do modelo, houve o pré-processamento dos dados, resultando em uma base de treinamento em que todos os atributos eram numéricos e não havia mais valores nulos.

4.3.1 Otimização de hiperparâmetros (Fase de treino)

Para este trabalho foi utilizado o modelo "*light gradient-boosting machine*", conhecido também como "*LightGBM*" (KE *et al.*, 2017). A obtenção de um modelo treinado passou por duas etapas de otimização, sendo que em cada etapa foram manipulados quatro de seus hiperparâmetros: profundidade máxima ("*max_depth*"), quantidade de folhas ("*num_leaves*"), taxa de aprendizagem ("*learning_rate*") e número de estimadores ("*n_estimators*").

A primeira etapa empregou uma estratégia de busca aleatória ("*random search*") sobre um espaço de busca maior, mais amplo, e a segunda etapa explorou em torno das coordenadas obtidas na primeira etapa através de uma busca em grade ("*grid search*").

Dentre as métricas existentes para avaliação da otimização dos hiperparâmetros, a ROC-AUC foi a utilizada nas duas etapas de busca (aleatória e em grade). A ROC-AUC é uma medida de desempenho usada para avaliar a capacidade discriminativa de modelos de classificação binária com base na curva ROC, variando de 0 a 1, valor que representa a área sob a curva ROC, também conhecida como ROC-AUC (*"area under the ROC curve"*). Uma pontuação ROC-AUC mais próxima de 1 indica um modelo de classificação mais competente, com uma alta capacidade de distinguir entre as classes positiva e negativa. Por outro lado, um valor próximo de 0,5 indica um modelo que tem um desempenho semelhante ao acaso, enquanto um valor próximo de 0 indica um modelo com classificações incorretas invertidas (BRADLEY, 1997).

Na busca aleatória, o espaço de busca foi dado dentro das combinações possíveis na lista de cada atributo, conforme a Tabela 7:

Tabela 7 - Espaço de busca utilizado na busca aleatória.

Atributo	Lista de valores
Profundidade máxima	3; 8; 13; 18; 23; 28; 33
Número de folhas	8; 13; 18; 23; 28; 33; 38; 43; 48; 53; 58; 63; 68
Taxa de aprendizagem	0,0001; 0,001; 0,01; 0,1
Número de estimadores	50; 150; 250; 350; 450; 550

Usando validação cruzada com 10 partes e usando como métrica de desempenho a ROC-AUC, obtivemos o seguinte conjunto de hiperparâmetros:

- Profundidade máxima: 3
- Número de folhas: 28
- Taxa de aprendizagem: 0,1
- Número de estimadores: 250

Após a busca aleatória, criou-se mais um espaço de busca, desta vez com valores no entorno do conjunto obtido anteriormente, procedendo-se a busca em grade. A Tabela 8 representa o espaço usado na busca em grade.

Tabela 8 - Espaço de busca utilizado na busca em grade.

Atributo	**Lista de valores**
Profundidade máxima	2; 3; 4; 5; 6
Número de folhas	26; 27; 28; 29; 30; 31
Taxa de aprendizagem	0,05; 0,075; 0,1; 0,125; 0,15
Número de estimadores	200; 225; 250; 275

Os hiperparâmetros obtidos após a otimização foram:

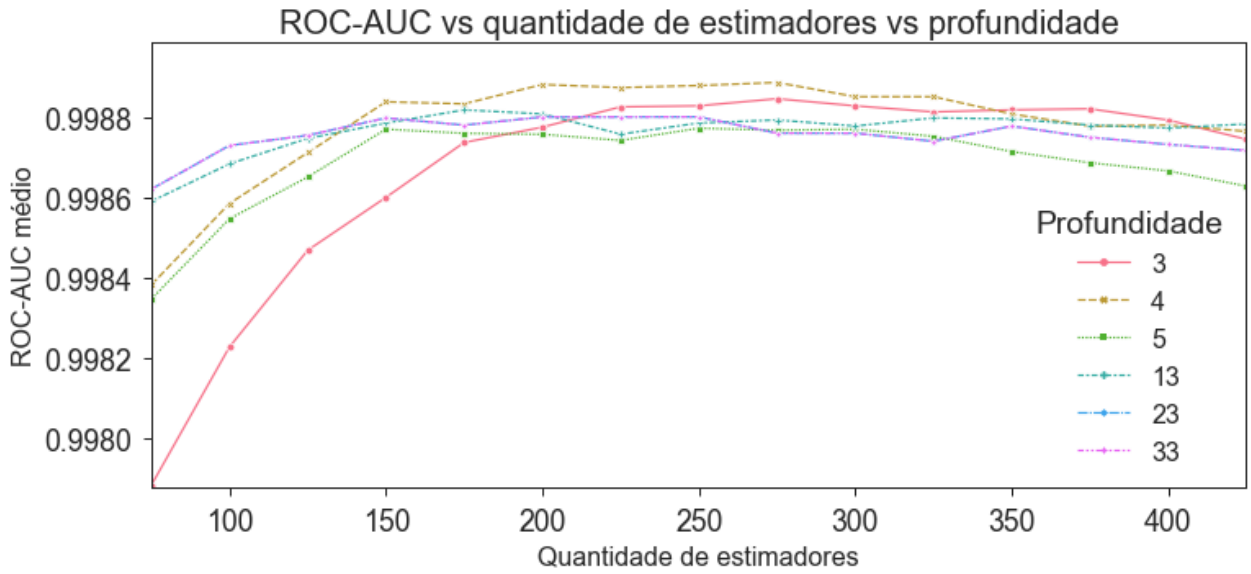
- Profundidade máxima: 4
- Número de folhas: 26
- Taxa de aprendizagem: 0,075
- Número de estimadores: 275

A Figura 8 evidencia que o melhor desempenho de ROC-AUC (0,99889) ocorre no ponto com 275 estimadores e profundidade 4, além de considerar o máximo de 26 folhas por estimador base e taxa de aprendizado de 0,075.

A configuração de máximo de 26 folhas por estimador base, taxa de aprendizado de 0,075 e profundidade de 4 foi superior, também, entre 200 e 325 estimadores.

Na Figura 8, estão sumarizados os resultados de ROC-AUC da otimização dos hiperparâmetros (busca aleatória e em grade).

Figura 8 - Evolução da métrica ROC-AUC de acordo com a quantidade de estimadores e profundidade usadas no espaço de busca.



Fonte: Autor.

4.3.2 Avaliação do desempenho do modelo (Fase de teste)

Após a otimização dos hiperparâmetros, o modelo foi testado com a base de teste contendo 702 endereços com atributos pré-processados, cerca de 15% da base total. As métricas utilizadas para a avaliação foram: acurácia, precisão, revocação (*recall*), F1-score e ROC-AUC, porém foi dada maior importância à acurácia e ao ROC-AUC, servindo de parâmetro de comparação com o trabalho de Farrugia; Ellul, Azzopardi (2020).

Para obtenção das métricas foram sorteados sem reposição 667 registros (95% da base de teste), em 10 sorteios, para a composição de média e desvio padrão das métricas de avaliação. Seguem os resultados:

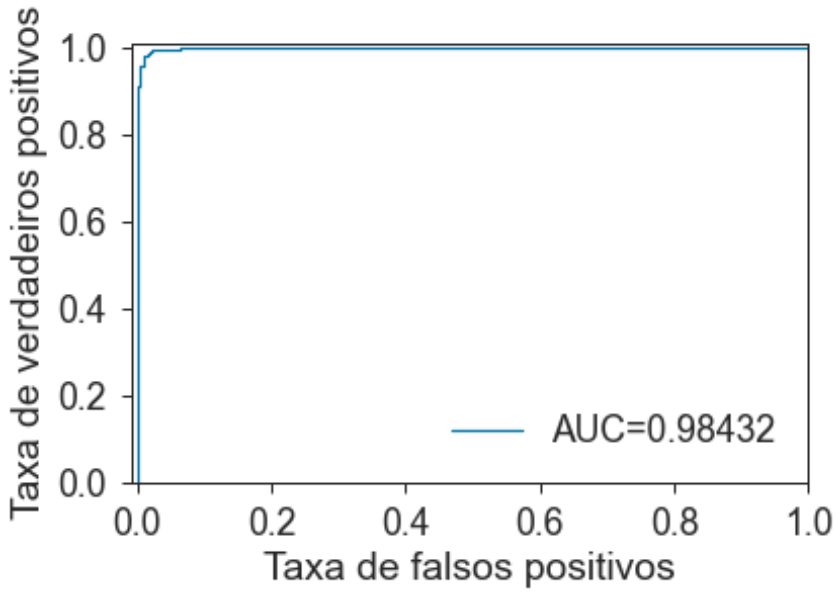
- Acurácia: 0,98438 ± 0,00105
- Precisão: 0,98173 ± 0,00232
- Revocação (*recall*): 0,98369 ± 0,00107
- F1-score: 0,98271 ± 0,00117
- ROC-AUC: 0,98432 ± 0.00098

Considerando uma validação contendo todos os 702 registros da base de teste, obtivemos a curva ROC da Figura 9 e a matriz de confusão ilustrada na Figura 10, além das métricas de desempenho abaixo.

- Acurácia: 0,98433
- Precisão: 0,98404
- Revocação (recall): 0,98432
- F1-score: 0,98417
- ROC-AUC: 0,98432

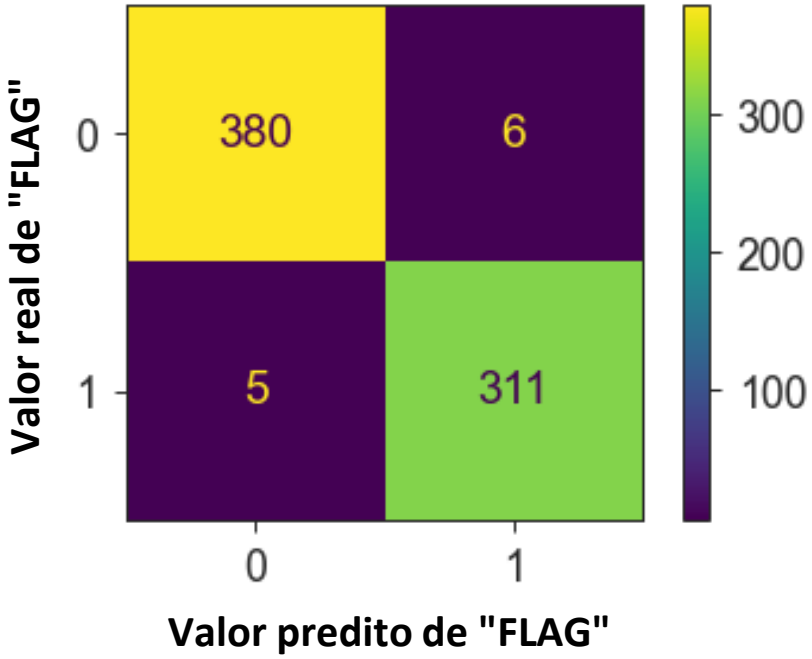
A área sob a curva ROC é obtida a partir do gráfico de taxa de verdadeiros positivos versus a taxa de falsos positivos. A AUC está indicada na Figura 9 é de 0,98432. Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1.

Figura 9 - Curva ROC e área sob a curva ROC.



Fonte: Autor.

Figura 10 - Matriz de confusão obtido com o conjunto de testes.



Fonte: Autor.

Considerando o conjunto de testes em sua totalidade (702 registros), ao aplicarmos o modelo 6 carteiras foram classificadas como falsos positivos e 5 carteiras, como falsos negativos.

4.4 Atributos mais importantes para a classificação

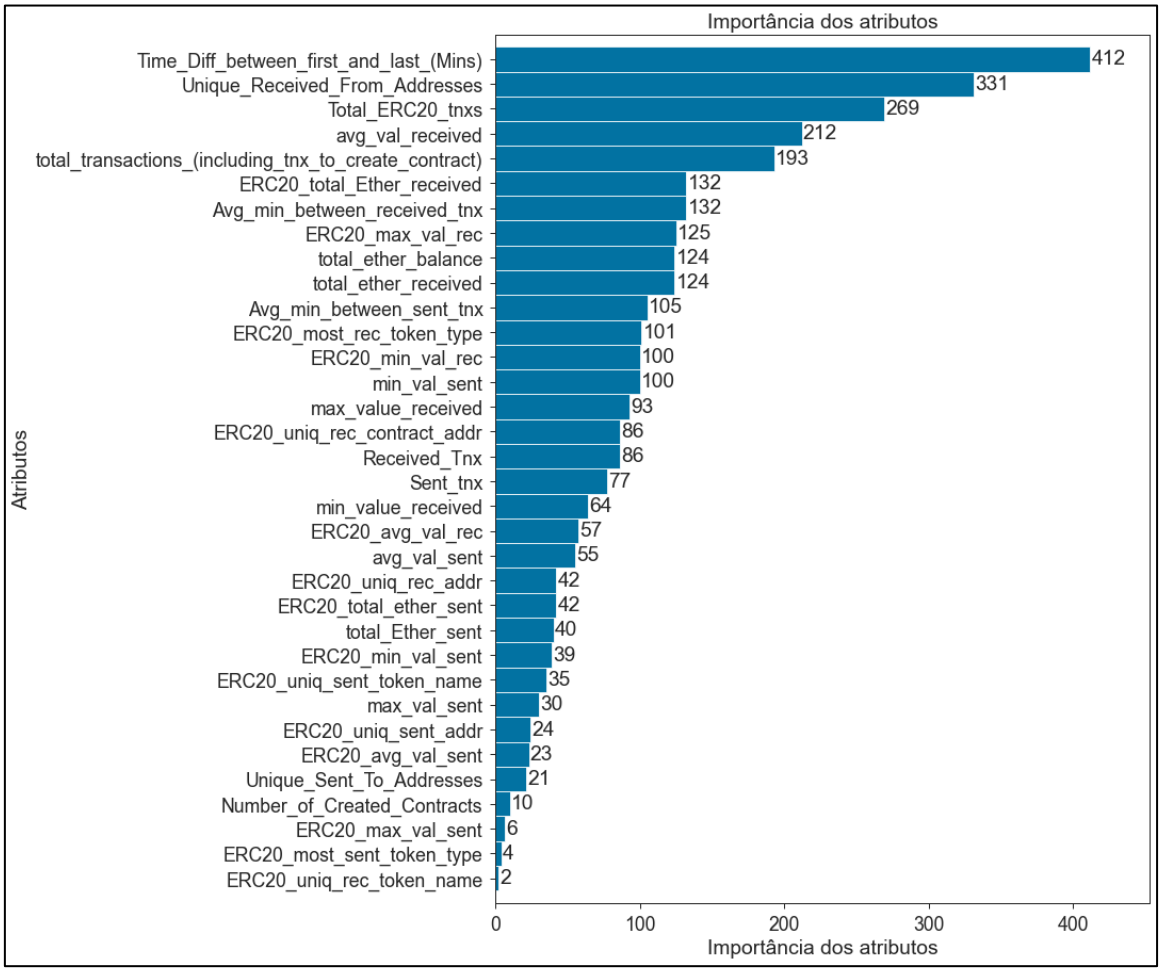
Através do modelo treinado, obtivemos a importância que os atributos tiveram para a classificação dos registros em lícitos e ilícitos, tendo por base dois critérios: a divisão de recursos e o ganho. A Figura 11 e Figura 12 mostram a importância de cada atributo sob a perspectiva dos dois critérios, respectivamente.

Motiva esclarecer que a divisão de recursos (*feature*) refere-se à medida de quão significativo cada atributo ou característica é para o modelo, em sua capacidade de tomar decisões de classificação. Ela pode ser obtida observando a contribuição de cada recurso nas divisões tomadas ao longo da construção do modelo (BREIMAN, 2001). Os recursos que desempenham um papel mais importante na classificação terão uma maior importância.

Tabela 9 - Os 10 atributos mais importantes (Divisão de recursos)

Atributos	Divisão de recursos
Time_Diff_between_first_and_last_(Mins)	412
Unique_Received_From_Addresses	331
Total_ERC20_tnxs	269
avg_val_received	212
total_transactions_(including_tnx_to_create_contract)	193
ERC20_total_Ether_received	132
Avg_min_between_received_tnx	132
ERC20_max_val_rec	125
total_ether_balance	124
total_ether_received	124

Figura 11 - Importância de divisão de recursos dos atributos.



Fonte: Autor.

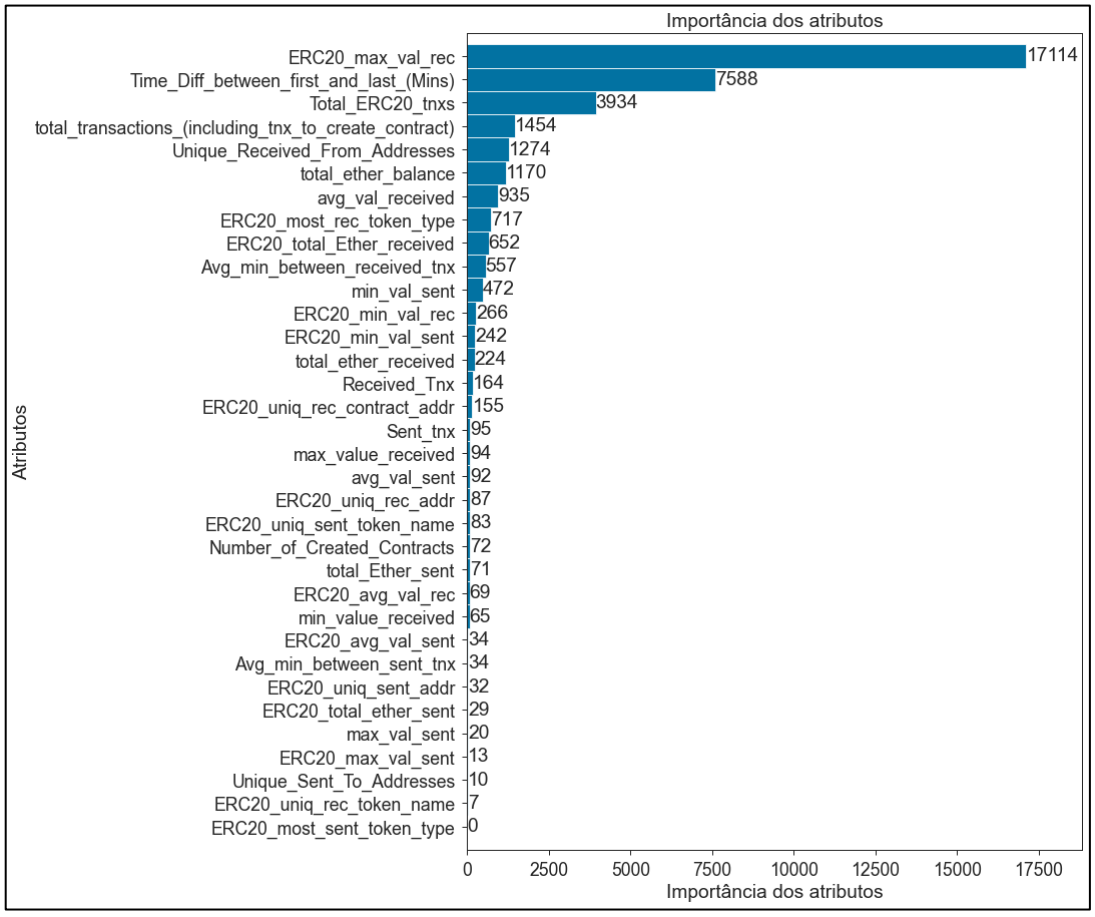
Por sua vez, o ganho (*gain*) está relacionado ao ganho de informação alcançado por cada recurso durante a construção do modelo. O ganho de informação é uma medida que quantifica

a redução da entropia ou impureza dos dados após uma divisão baseada em um determinado recurso. O ganho de informação é calculado e comparado para cada recurso, e aquele que proporcionar o maior ganho é considerado mais importante (FREUND; SCHAPIRE, 1997).

Tabela 10 - Os 10 atributos mais importantes (Ganho)

Atributos	Ganho
ERC20_max_val_rec	17114,4
Time_Diff_between_first_and_last_(Mins)	7588,4
Total_ERC20_tnxs	3934,3
total_transactions_(including_tnx_to_create_contract)	1453,9
Unique_Received_From_Addresses	1274,3
total_ether_balance	1169,7
avg_val_received	934,8
ERC20_most_rec_token_type	717,3
ERC20_total_Ether_received	652,4
Avg_min_between_received_tnx	557,1

Figura 12 - Importância de ganho dos atributos.

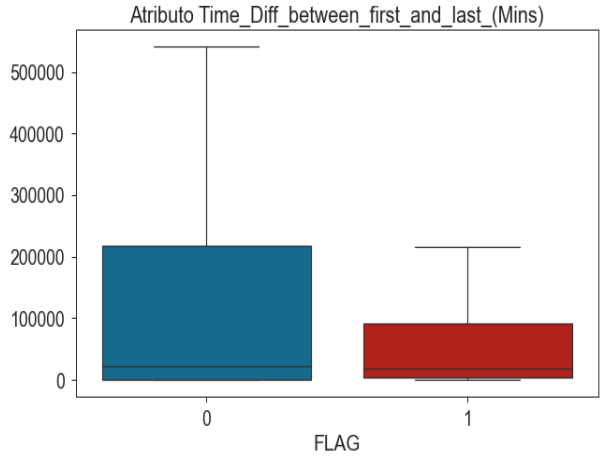


Fonte: Autor.

Aqui cabe ressaltar que os achados corroboram o trabalho de Farrugia, Ellul e Azzopardi (2020), no qual o atributo “Time diff_between first and last (Mins)” foi um dos mais impactantes no resultado do modelo utilizado naquele trabalho, igualmente à nossa informação encontrada. Esse fato indica que nos casos de suspeitas de carteiras ilícitas, a diferença em minutos, entre a primeira e a última transação em uma carteira é um fator importante para a sua caracterização e de atenção para as autoridades judiciais.

Como pode ser demonstrado na figura de *boxplot* a seguir (Figura 13), o bloco azul (carteira lícita) possui um intervalo maior em minutos entre a primeira e a última transação, quando comparado com o bloco vermelho (carteira ilícita). Esse curto intervalo de tempo pode ser interpretado como comportamento semelhante ao que se observa em processos de lavagem de dinheiro, em que contas de passagem (*pooling accounts*) são utilizadas para inúmeras transferências entre contas, ocasião em que o valor não fica depositado por muito tempo (pouco tempo de utilização, servindo a com meramente como “intermediária”), conforme apontou Lima (2021).

Figura 13 - Atributo “Time diff_between first and last (Mins)”

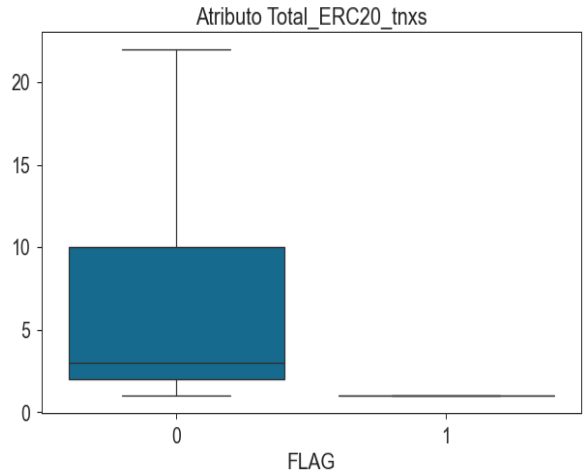


Fonte: Autor.

Em complemento à análise temporal, pode-se investigar o aspecto de quantidade ao analisar o atributo “Total_ERC20_tnxs” (Número total de transações de transferência de token ERC20), destacando-se a diferença na distribuição dos valores entre as carteiras lícitas e ilícitas, conforme Figura 14. A baixa quantidade de transações dos endereços tidos como ilícitos pode sinalizar que as contas são utilizadas uma única vez ou poucas vezes, e descartadas em seguida, o que permite especular ser uma ação para evitar o monitoramento pelos órgãos de controle e

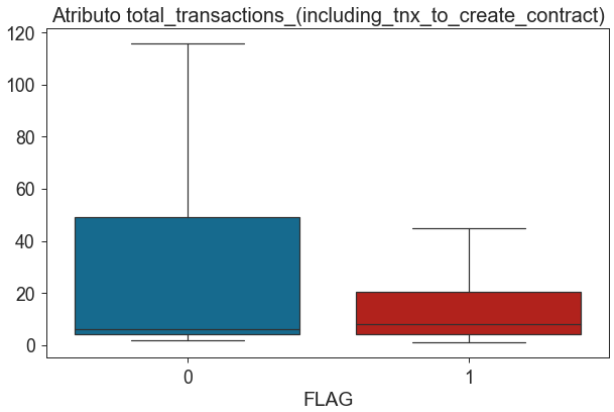
fiscalização, o que vai ao encontro do que seria a fase ocultação (*layering*) da lavagem de dinheiro (FATF, 2005). Esse comportamento é corroborado quando se observa a importância para o modelo classificador do atributo "total_transactions_(including_tnx_ to_create_contract)", o qual se refere ao número total de transações efetuadas na conta, sendo que as contas ilícitas apresentam um número reduzido de transações em comparação às lícitas (Figura 15).

Figura 14 - Atributo "Total_ERC20_tnx"



Fonte: Autor.

Figura 15 - Atributo "total_transactions_(including_tnx_ to_create_contract)"

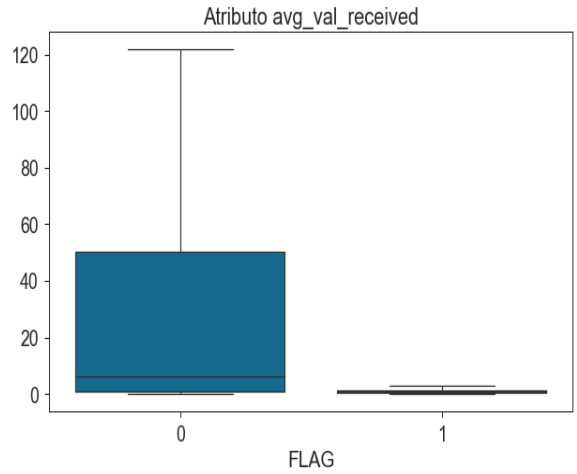


Fonte: Autor.

Nessa esteira, examinando aspectos de volume financeiro, apresentam-se boxplots para os atributos "avg_val_received" (Valor médio em Ether recebido) e "ERC20_max_val_rec" (Valor máximo em Ether recebido de transações de token ERC20), novamente com grande diferença na distribuição dos valores entre as carteiras analisadas. Esse achado pode indicar o uso da técnica conhecida por *smurfing*, que é uma prática comum

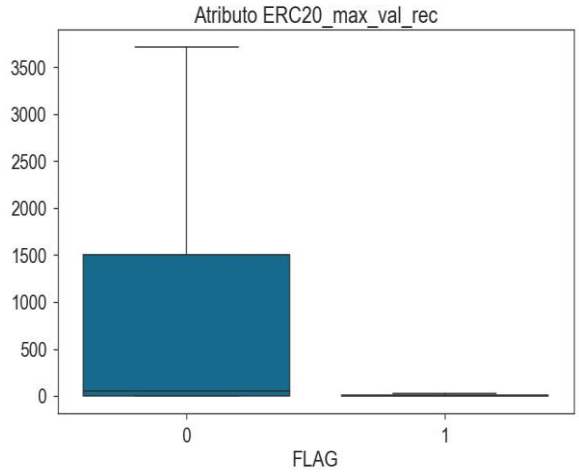
relacionada ao crime de lavagem de dinheiro, na qual o intuito é dividir altos valores em transações menores e de baixos valores, conforme apontam Irwin *et al.* (2011) e EIFREM (2019), o que pode ser observado nas Figura 16 e Figura 17.

Figura 16 - Atributo “avg_val_received”



Fonte: Autor.

Figura 17 - Atributo “ERC20_max_val_rec”



Fonte: Autor.

5 CONCLUSÃO

A tecnologia blockchain, com destaque para a plataforma Ethereum, revolucionou diversos setores ao possibilitar a criação de contratos inteligentes, aplicativos descentralizados e organizações autônomas descentralizadas. No entanto, esse avanço tecnológico também despertou o interesse de criminosos, que encontraram na utilização de criptomoedas uma oportunidade para a prática da lavagem de dinheiro. A baixa regulamentação e a natureza pseudoanônima das transações tornam a identificação de contas ilícitas um desafio significativo. Nesse contexto, o presente estudo teve por objetivo investigar em que medida é possível identificar endereços de criptomoedas suspeitos de envolvimento com a prática da lavagem de dinheiro, por meio do emprego de técnicas de aprendizado de máquina.

O método de pesquisa desta dissertação consistiu em aprofundar a investigação sobre a identificação de endereços de criptomoedas suspeitos de envolvimento com a lavagem de dinheiro, inspirado na pesquisa de Farrugia, Ellul, Azzopardi (2020), pesquisadores que utilizaram ferramentas de *machine learning* (modelo XGBoost) no contexto da blockchain Ethereum, para detectar padrões em endereços na rede associados a atividades ilícitas.

Alternativamente, o modelo aqui proposto foi o LightGBM (*Light Gradient Boosting Machine*), difundido por KE *et al.* (2017), que é um algoritmo de aprendizado de máquina que utiliza a técnica de *boosting* para criar modelos de classificação e regressão, apresentando uma abordagem chamada "*histogram-based*" que visa otimizar o processo de aprendizado, reduzindo a complexidade computacional e melhorando a eficiência do treinamento.

Partindo da mesma base de dados, o desempenho apresentado pelo modelo LightGBM foi semelhante ao modelo XGBoost utilizado por Farrugia, Ellul e Azzopardi (2020), tendo por base às duas principais métricas adotadas: acurácia e ROC-AUC. Outrossim, os atributos mais importantes na classificação de carteiras ilícitas foram, de certa forma, corroborados com os resultados encontrados no trabalho base. A utilização do LightGBM neste trabalho apresentou a métrica Acurácia de 0,98433, superior ao trabalho base de Farrugia, Ellul e Azzopardi (2020) que apresentou acurácia média 0,963 (±0,006).

Cabe enfatizar que o atributo mais impactante no modelo de classificação, considerando a métrica da divisão de recursos (*feature)*, foi o “Time diff_between_first_and_last_(Mins)”, ou seja, a diferença em minutos entre a primeira e a última transação em uma carteira é um fator importante na identificação de endereços suspeitos. Esse curto período de “vida” pode indicar a utilização de contas de passagem (*pooling accounts*), em que inúmeras transferências são feitas em pouco tempo.

Em relação aos aspectos de quantidade de operações e volume das transações, os atributos mais relevantes foram “Total_ERC20_tnxs” (Número total de transações de transferência de token ERC20), “total_transactions (including_tnx_ to_create_contract)” (número total de transações efetuadas na conta), “avg_val_received” (Valor médio em Ether recebido) e “ERC20_max_val_rec” (Valor máximo em Ether recebido de transações de token ERC20). A análise conjunta desses atributos pode levar a interpretação de contas que são utilizadas poucas vezes e descartadas em seguida, o que pode suscitar ações que buscam minimizar exposição junto a órgãos de controle e fiscalização, o que vai ao encontro do que seria a fase ocultação (*layering*) da lavagem de dinheiro, além de sinalizar pela prática do *smurfing* (fracionamento de altos valores).

Diante da complexidade e atualidade do tema, acredita-se que a pesquisa possui grande relevância acadêmica e prática, fornecendo percepções valiosas, tanto para as instituições de monitoramento, como as de prevenção e combate à lavagem de dinheiro. No contexto nacional, a abordagem deste trabalho e de outras iniciativas na área acadêmica podem ser utilizadas pelas instituições no sentido de atender às novas legislações sobre o tema, com o intuito de monitorar e fiscalizar o mercado de criptomoedas e ativos virtuais.

Nas questões de limitações para o trabalho, as maiores dificuldades foram em relação à base de dados, pois para se treinar o modelo é preciso de uma base balanceada entre carteira lícitas e ilícitas. No caso das carteiras ilícitas, sua informação é dada por denúncias ou o envolvimento com carteiras de golpe ou fraude, assim a formação dessa base depende em sua maior parte de queixas ou envolvimento em atos pretéritos de fraude. Além disso, são poucas as bases de dados disponíveis ao público com essas características, o que leva a disponibilização da base de dados deste trabalho no repositório Github, o que pode fomentar futuros trabalhos.

Nesse sentido, existe um grande leque de possibilidades. A utilização de aprendizado de máquina no combate à lavagem de dinheiro se mostra oportuno, tanto em outras criptomoedas como no sistema financeiro (bancário) tradicional, dependendo muito da prospecção de bancos de dados confiáveis para utilização. A lavagem de dinheiro é um mal que está presente na maior parte das atividades criminosas de grande vulto, contaminando o sistema financeiro legal, e é necessário o seu combate, independente do meio empregado, e os modelos de aprendizado de máquina são uma ferramenta nova e poderosa no combate a esse tipo de crime.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AKARTUNA, Eray Arda; JOHNSON, Shane D.; THORNTON, Amy E. **The money laundering and terrorist financing risks of new and disruptive technologies: a futures-oriented scoping review**. [*S. l.*]: Palgrave Macmillan UK, 2022-. ISSN 17434645. Disponível em: https://doi.org/10.1057/s41284-022-00356-z.

ALARAB, Ismail; PRAKOONWIT, Simant; NACER, Mohamed Ikbal. Comparative Analysis Using Supervised Learning Methods for Anti-Money Laundering in Bitcoin. **ACM International Conference Proceeding Series**, [*s. l.*], p. 11–17, 2020.

AZIZ, Rabia Musheer *et al.* LGBM: a machine learning approach for Ethereum fraud detection. **International Journal of Information Technology (Singapore)**, [*s. l.*], v. 14, n. 7, p. 3321–3331, 2022.

AZIZ, Rabia Musheer *et al.* Modified Genetic Algorithm with Deep Learning for Fraud Transactions of Ethereum Smart Contract. **Applied Sciences (Switzerland)**, [*s. l.*], v. 13, n. 2, 2023.

BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, [*s. l.*], v. 30, n. 7, p. 1145–1159, 1997.

BREIMAN, Leo. Random Forests. [*s. l.*], p. 542–545, 2001.

BUTERIN, Vitalik Vitalin. Ethereum: inteligentny kontrakt nowej generacji i zdecentralizowana platforma aplikacji. **Whitepaper**, [*s. l.*], n. January, p. 1–36, 2014.

CAMPBELL-VERDUYN, Malcolm. Laundering Governance. **Crime Law Soc Change**, [*s. l.*], v. 69, p. 283–305, 2018.

CHAINALYSIS INC. The 2023 Crypto Crime Report. [*s. l.*], n. February, 2023a. Disponível em: https://go.chainalysis.com/rs/503-FAP-074/images/2020-Crypto-Crime-Report.pdf.

CHAVES, Alan Rodrigues. **Caracterizando a evolução de software de contratos inteligentes: um estudo exploratório-descritivo utilizando github e etherscan**. 2021. - UNIVERSIDADE FEDERAL DO CEARÁ CAMPUS DE CRATEÚS, [*s. l.*], 2021.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A scalable tree boosting system. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, [*s. l.*], v. 13-17-Augu, p. 785–794, 2016.

CONNOLLY, Lena Y.; WALL, David S. The rise of crypto-ransomware in a changing cybercrime landscape: Taxonomising countermeasures. **Computers and Security**, [*s. l.*], v. 87, 2019.

DOMINGOS, Pedro. A Few Useful Things to Know About Machine Learning. **Communications of the ACM**, [*s. l.*], v. 55, n. 10, 2012. Disponível em: https://dl.acm.org/citation.cfm?id=2347755.

EIFREM, Emil. How graph technology can map patterns to mitigate money-laundering risk.

Computer Fraud and Security, [*s. l.*], v. 2019, n. 10, p. 6–8, 2019. Disponível em: http://dx.doi.org/10.1016/S1361-3723(19)30105-8.

FANUSIE, Yaya J; ROBINSON, Tom. Bitcoin Laundering: An Analysis of Illicit Flows into Digital Currency Services. **Center on Sanctions and Illicit Finance …**, [*s. l.*], p. 15, 2018. Disponível em: https://www.defenddemocracy.org/content/uploads/documents/MEMO_Bitcoin_Laundering.pdf.

FARRUGIA, Steven; ELLUL, Joshua; AZZOPARDI, George. Detection of illicit accounts over the Ethereum blockchain. **Expert Systems with Applications**, [*s. l.*], v. 150, n. February 2019, p. 113318, 2020. Disponível em: https://doi.org/10.1016/j.eswa.2020.113318.

FATF. Money Laundering and Terrorist Financing Typologies. **World Trade**, [*s. l.*], v. 268, n. June, p. 44, 2005. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/7754383.

FREUND, Yoav; SCHAPIRE, Robert E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, [*s. l.*], v. 55, n. 1, p. 119–139, 1997.

HUANG, K. An Optimized LightGBM Model for Fraud Detection. **Journal of Physics: Conference Series**, [*s. l.*], v. 1651, n. 1, 2020.

IBRAHIM, Rahmeh Fawaz; ELIAN, Aseel Mohammad; ABABNEH, Mohammed. Illicit Account Detection in the Ethereum Blockchain Using Machine Learning. **2021 International Conference on Information Technology, ICIT 2021 - Proceedings**, [*s. l.*], p. 488–493, 2021.

IRWIN, Angela Samantha Maitland; CHOO, Kim-Kwang Raymond; LIU, Lin. An analysis of money laundering and terrorism financing typologies. **Journal of Money Laundering Control**, [*s. l.*], v. 15, n. 1, p. 85–111, 2011.

JOSEPH, V. Roshan. Optimal ratio for data splitting. **Statistical Analysis and Data Mining**, [*s. l.*], v. 15, n. 4, p. 531–538, 2022.

JULLUM, Martin *et al.* Detecting money laundering transactions with machine learning. **Journal of Money Laundering Control**, [*s. l.*], v. 23, n. 1, p. 173–186, 2020.

KADOO, Shubham; SODI, Khushboo. An Analysis of Cryptocurrency, Bitcoin and the Future. **International Journal of Advanced Research in Science, Communication and Technology**, [*s. l.*], v. 3, n. 3, p. 287–292, 2023.

KE, Guolin *et al.* LightGBM: A highly efficient gradient boosting decision tree. **Advances in Neural Information Processing Systems**, [*s. l.*], v. 2017-Decem, n. Nips, p. 3147–3155, 2017.

KRUISBERGEN, E. W. *et al.* Money talks money laundering choices of organized crime offenders in a digital age. **Journal of Crime and Justice**, [*s. l.*], v. 42, n. 5, p. 569–581, 2019.

LARSEN, Jan; GOUTTE, Cyril. On Optimal Data Split for Generalization Estimation and

Model Selection. **Design**, [*s. l.*], 1999.

LIMA, Rafael Sousa. **Análise de Redes Sociais no Combate aos Crimes de Lavagem de Dinheiro e Corrupção**. 2021. - Universidade de Brasília - UnB, [*s. l.*], 2021.
LITTLE, Roderick J. A.; RUBIN, Donald B. **Statistical Analysis with Missing Data**. 3ª Ediçãoed. [*S. l.: s. n.*], 2020-. ISSN 09641998.

LORENZ, Joana *et al.* Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity. **ICAIF 2020 - 1st ACM International Conference on AI in Finance**, [*s. l.*], 2020.

M. JOHN, Linda *et al.* Predicting House Prices using Machine Learning and LightGBM. **SSRN Electronic Journal**, [*s. l.*], 2022.

MICHAEL, Epimack *et al.* An Optimized Framework for Breast Cancer Classification Using Machine Learning. **BioMed Research International**, [*s. l.*], v. 2022, 2022.

MÖSER, Malte; BÖHME, Rainer; BREUKER, Dominic. An Inquiry into Money Laundering Tools. **2013 APWG eCrime Researchers Summit**, [*s. l.*], p. 1–14, 2013a. Disponível em: https://informationsecurity.uibk.ac.at/pdfs/MBB2013-ECRIME.pdf.

NAKAMOTO, Satoshi. Bitcoin: A Peer-to-Peer Electronic Cash System. **Bitcoin.Org**, [*s. l.*], p. 1–9, 2008a. Disponível em: https://bitcoin.org/files/bitcoin-paper/bitcoin_pt_br.pdf.
NAKAMOTO, Satoshi. **Bitcoin: A Peer-to-Peer Electronic Cash System**. [*S. l.: s. n.*], 2008b.

NICA, Octavian; PIOTROWSKA, Karolina; SCHENK-HOPPP, Klaus Reiner. Cryptocurrencies: Economic Benefits and Risks. **SSRN Electronic Journal**, [*s. l.*], p. 1–56, 2017.

PATTERSON, Josh; GIBSON, Adam. **Deep Learning - A Practitioner's Approach**. Sebastopol: O'Reilly Media, Inc., 2017. v. First Edit

PILKINGTON, Marc. Blockchain Technology and Its Applications. **Research Handbook on Digital Transformations, edited by F. Xavier Olleros and Majlinda Zhegu. Edward Elgar, 2016, Available at SSRN: https://ssrn.com/abstract=2662660**, [*s. l.*], p. 1247–1260, 2015.

RUIZ, Eric Pettersson; ANGELIS, Jannis. Combating money laundering with machine learning – applicability of supervised-learning algorithms at cryptocurrency exchanges. **Journal of Money Laundering Control**, [*s. l.*], v. 25, n. 4, p. 766–778, 2022.

TELLES, Christiana Mariani da Silva. Sistema bitcoin, lavagem de dinheiro e regulação. **Dissertação de Mestrado - FGV**, [*s. l.*], p. 144, 2018a.

TELLES, Christiana Mariani da Silva. **Sistema bitcoin, lavagem de dinheiro e regulação**. 2018b. 144 f. [*s. l.*], 2018. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27350/DISSERTACAO-FINAL-13fev19-Christiana M S Telles.pdf?sequence=1&isAllowed=y.

TIWARI, Milind; GEPP, Adrian; KUMAR, Kuldeep. A review of money laundering literature: the state of research in key areas. **Pacific Accounting Review**, [*s. l.*], v. 32, n. 2, p. 271–303, 2020.

TREVOR, Hastie; ROBERT, Tibshirani; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2ª Ediçãoed. [*S. l.: s. n.*], 2013.

VAN WEGBERG, Rolf; OERLEMANS, Jan Jaap; VAN DEVENTER, Oskar. Bitcoin money laundering: mixed results?: An explorative study on money laundering of cybercrime proceeds using bitcoin. **Journal of Financial Crime**, [*s. l.*], v. 25, n. 2, p. 419–435, 2018.

ZOU, Weiqin *et al.* Smart Contract Development : Challenges and Opportunities. **IEEE Transactions on Software Engineering**, [*s. l.*], v. 47, n. 10, p. 2084–2106, 2021.