Classificação de Coleções de NFTs Explorando Metadados e Aprendizagem de Máquina*

Samuel de Oliveira Ribeiro[1], Dayan Ramos Gomes[1], Nara Raquel D. Andrade[1]
Emanuel Aurélio F. de Miranda[1], Glauber Dias Gonçalves[1]

[1]Universidade Federal do Piauí (UFPI) - CSHNB

{samuel, dayan, nara, emanuel, ggoncalves}@ufpi.edu.br

Abstract. *Non-fungible Tokens (NFTs) are digital objects with unique identities and ownership verified via blockchain networks. The digital arts and media industry has embraced NFTs for their secure features, such as defining authorship, transfer, and royalties, which can be programmed into smart contracts. The classification of NFTs is crucial for their commercialization but often relies on the author's definition, which may be prone to errors, or on expert evaluation. In this work, we analyze NFT collection classes based on metadata extracted from OpenSea, the largest NFT platform. We assess the efficiency of supervised machine learning to identify the most relevant attributes of these collections and classify them into the nine most popular categories on the platform. Our results demonstrate the challenges of automating classification to assist users and platform curators, achieving a promising accuracy of 67% and an F1 score of 72% in the best cases.*

1. Introdução

Token não fungível, ou NFT, do inglês *Non-fungible Token*, é um objeto digital registrado em plataformas blockchain, tipicamente associado a conteúdo de texto ou imagem, que lhe confere características únicas, tornando-o também um objeto colecionável. Adicionalmente, o NFT permite a definição de um autor (criador), transferências de propriedade entre usuários, *royalties* para o criador, entre outros recursos do ambiente distribuído de blockchains. Devido a essas características, NFTs vêm sendo adotados por artistas para a criação e distribuição de conteúdo digital, visando a proteção do direito autoral e o ganho com *royalties* na revenda de itens [Wang et al. 2021]. A plataforma Ethereum é pioneira na emissão de NFTs e concentra a maioria das coleções. Ela oferece padrões de programação específicos para NFTs, onde cada um é único, com atributos específicos. Esses atributos incluem, especialmente, um link para um arquivo de mídia temático (e.g., jogos, artes, vídeos ou redes sociais), identificadores do autor e proprietário, valores de venda e *royalties* para o autor. As transferências de propriedade são registradas na blockchain, garantindo descentralização, transparência e imutabilidade nas transações, corroborando a credibilidade no comércio de NFTs.

Nesse contexto, NFTs abrem um novo caminho para utilização e veiculação de objetos digitais, i.e., *tokens*, sob a Internet, onde o aspecto mais relevante é a autoria ou propriedade do *token*. A classificação de NFTs é uma tarefa importante para a sua

*Esta pesquisa é financiada com o apoio da Fundação de Amparo à Pesquisa do Piauí (FAPEPI) processo no. 00110.000235/2022-78 e Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI) UFPI.

comercialização mas dependente da definição do próprio autor, que pode se equivocar por inexperiência na área, ou da avaliação de um especialista. Neste trabalho, analisamos classes de coleções de NFTs baseado em metadados dessas coleções extraídos do *OpenSea*, a maior plataforma de NFTs na atualidade. Avaliamos a eficiência de aprendizagem de máquina supervisionada para aprender os atributos mais relevantes dessas coleções e classificá-las nas 9 categorias mais populares nessa plataforma. Mostramos o quão desafiante é automatizar a classificação para auxiliar usuários e curadores da plataforma.

Em suma este trabalho visa as seguintes contribuições: (1) identificação de metadados, i.e., atributos descritivos, assim como estatísticas de comercialização das coleções de NFTs importantes para a sua classificação e curadoria; e (2) um método automático para classificar coleções de NFT baseado nos atributos de metadados acima referidos.

2. Trabalhos Relacionados

Existem vários trabalhos que utilizam técnicas de aprendizagem de máquina com atributos extraídos de transações na blockchain para a classificação de perfis de usuários, seja para fins comerciais ou para a segurança da rede. Em [Valadares et al. 2023], são propostas técnicas para a classificação de usuários como comuns ou profissionais com base nas transações realizadas na plataforma Ethereum, extraídas a partir da API Etherscan. Em [Aspembitova et al. 2021], é analisado o comportamento dos usuários no mercado de criptoativos, focando na estrutura comportamental dos usuários que realizam transações de compra e venda de criptoativos. Em [Xu et al. 2020], foi desenvolvido um modelo de classificação com *random forest* para diferenciar o tráfego malicioso em um tipo de "ataque de eclipse", onde um ator mal-intencionado pode interferir com os nós de uma rede. Uma rede neural LSTM foi desenvolvida por [Hu et al. 2021] para identificar os tipos mais comuns de contratos inteligentes no Ethereum.

Por sua vez, o crescimento na negociação de NFTs também vem atraindo pesquisas com aplicações de aprendizagem de máquina e mineração de dados. Em [Casale-Brunet et al. 2021], foi feita uma análise do comportamento dos usuários que comercializavam as oito coleções de NFTs mais populares da plataforma *OpenSea*. O estudo de [Costa et al. 2023] propõe um método para prever o preço de venda de NFTs utilizando aprendizagem de representação multimodal, ou seja, análise de conteúdo visual do NFT e também sua descrição textual. Já [Nadini et al. 2021] propõe uma análise mais generalista da rede para entender como o mercado funciona e prever vendas. Foram conduzidas análises utilizando dados das propriedades estatísticas, rede de compradores e vendedores e as características visuais dos itens. Contudo, nenhum desses trabalhos foca em identificar atributos relevantes de coleções de NFTs que permitam sua classificação automática, facilitando o trabalho de curadoria das plataformas e dos usuários desse ecossistema.

3. Metodologia

No presente estudo, coletamos metadados de coleções de NFTs a partir da plataforma *OpenSea*, uma das maiores e mais populares plataformas de comercialização de NFTs no momento da escrita deste trabalho. Esta escolha se deve ao suporte tecnológico robusto que a plataforma oferece para análise de dados, facilitando a obtenção de informações relevantes. A *OpenSea* disponibiliza uma API REST que permite acessar metadados das

Tabela 1. Categorias de coleções de NFTs utilizados na *OpenSea*.

Categoria	N° Coleções	Descrição
Art	247507 (57.86%)	Obra de arte cunhada em Blockchain, podem ser obras de arte físicas digitalizadas ou podem ser criadas nativamente usando ferramentas digitais.
PFPs	45298 (10.59%)	*Profile pictures* são itens digitais que representam a propriedade de uma imagem ou obra de arte única e colecionável que pode ser usada como foto de perfil.
Memberships	17931 (4.19%)	Tipo de *token* não fungível que fornece acesso a uma experiência, utilidade ou comunidade e, em alguns casos, todos os três.
Music	35758 (8.36%)	Utilizado como um certificado digital de propriedade que usa blockchain para verificar e proteger a propriedade de um conteúdo relacionado à música.
Photography	31406 (7.34%)	Item digital que representa uma fotografia digital ou imagem em movimento.
Gaming	21766 (5.09%)	NFTs associados a qualquer item digital do reino dos jogos virtuais e do metaverso como personagens, *skins*, personalizações, mapas, itens colecionáveis - qualquer criação digital que alguém usaria em ambientes de jogos virtuais.
Virtual Worlds	20696 (4.84%)	Ambientes digitais dentro de mundos virtuais ou metaversos, representando uma parte do espaço virtual que os usuários podem possuir, desenvolver e interagir.
Domain Names	3119 (0.73%)	Um *domain name* NFT é um *token* que armazena um nome de domínio no blockchain em vez de no tradicional Sistema de Nomes de Domínio, ou DNS.
Sports	4307 (1.01%)	Coleções de NFTs relacionados a esportes, como cartões colecionáveis digitais, arte esportiva e momentos históricos em formato digital.

coleções comercializadas, tornando o processo de coleta de dados eficiente e estruturado para fins de pesquisa e análise.[1] Esses metadados são informações preenchidas pelos autores sobre a coleção no momento da sua criação, assim como informações alimentadas pela própria plataforma com estatísticas sobre a comercialização da coleção.

A API *OpenSea* oferece uma variedade de pontos de acesso via API REST (*endpoints*), cada um fornecendo dados específicos sobre as coleções da plataforma, como preços, histórico de transações e informações de propriedade. Desenvolvemos um coletor em Python 3 para realizar requisições HTTP aos *endpoints* gratuitos da API REST *OpenSea*. Utilizamos *endpoints* das versões 1 e 2 da API; entretanto, em 01/2024, a versão 1 foi descontinuada, permanecendo apenas a versão 2. Os *endpoints* usados foram */api/v1/collection/collection-name* (*endpoint*-1), que solicita dados de uma coleção individual, e */api/v2/collections?chain=ethereum&limit=limit&next=next* (*endpoint*-2), que fornece uma lista de identificadores (slug) de todas as coleções. Ressaltamos que a API gratuita aplica limitações a requisições sem uma chave de acesso. Para este trabalho, solicitamos uma chave de acesso para fins de pesquisa, permitindo requisições de alto volume de dados.

Primeiramente, realizamos requisições no *endpoint*-2 para obter identificadores do maior número possível de coleções na plataforma. Em seguida, para cada coleção identificada, fizemos requisições no *endpoint*-1 para coletar seus metadados. Ao todo, coletamos 7.182.800 coleções, cada uma armazenada em um arquivo formato *JSON*, conforme o padrão da API *OpenSea*. Essas requisições foram realizadas entre 01/10/2023 e 31/11/2023, de modo que as coleções e seus metadados referem-se a esse período. Após a coleta, processamos os metadados. Inicialmente, selecionamos apenas coleções com o atributo *category* definido em seu JSON, dado o foco do trabalho na categorização de coleções. Assim, das coleções coletadas, 427.788 possuíam este atributo. A *OpenSea* define nove categorias distintas, conforme descrito na Tabela 1. A categoria é escolhida pelo autor da coleção no ato de sua criação, a partir das opções oferecidas pela plataforma. Com os metadados das coleções de NFT coletados em formato JSON, procedemos ao processamento desses dados. A Tabela 2 detalha os metadados extraídos.

[1]https://docs.opensea.io/reference/api-overview

Tabela 2. Atributos contidos em metadados de coleções de NFTs extraídos da plataforma *OpenSea*

Atributo	Descrição
total-volume	Total de vendas no *OpenSea* em ETH.
total-sales	Número total de transações de venda que ocorreram dentro da coleção.
total-supply	Total de NFTs criados para uma coleção.
num-owners	Número de proprietários de NFTs distintos na coleção.
average-price	Preço médio de venda dos itens da coleção no *OpenSea*.
num-reports	Número de relatórios de abuso no *OpenSea* sobre NFTs da coleção.
market-cap	Capitalização de mercado da coleção no *OpenSea*.
qtd-traits	Número de características visuais de todos os NFTs da coleção.
floor-price	Preço mínimo de venda dos NFTs da coleção no *OpenSea*
qtd-editors	Quantidade de editores da coleção.
category	Categoria da coleção utilizada no OpenSea

O próximo passo do processamento de dados foi realizar a análise estatística das características de todas as coleções. Essas análises destacam a variabilidade dos valores e medidas de centralidade, proporcionando uma compreensão abrangente das características das coleções de NFTs selecionadas. A Figura 1 apresenta o valor médio e

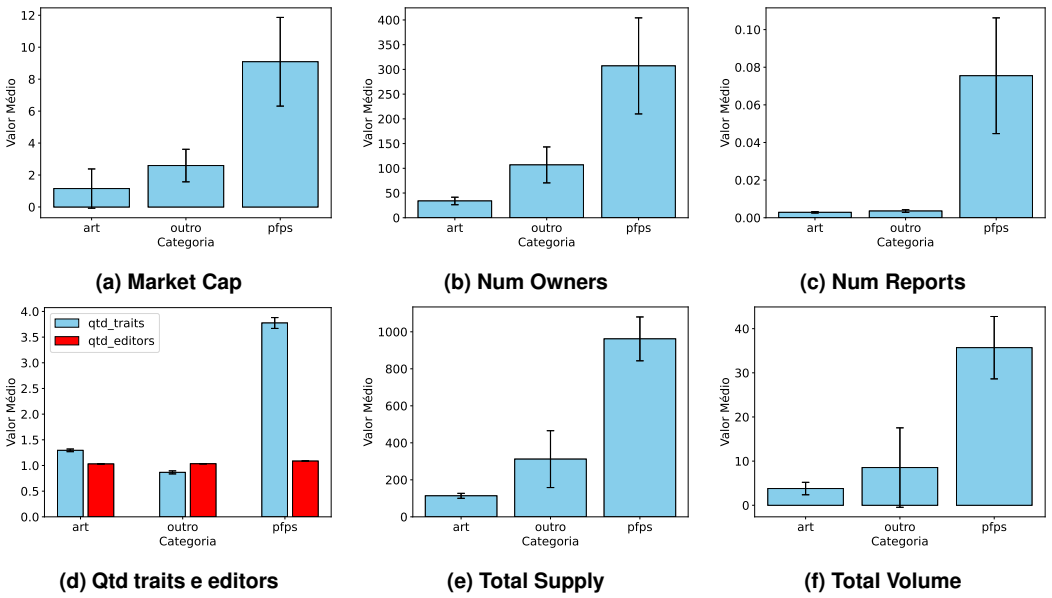
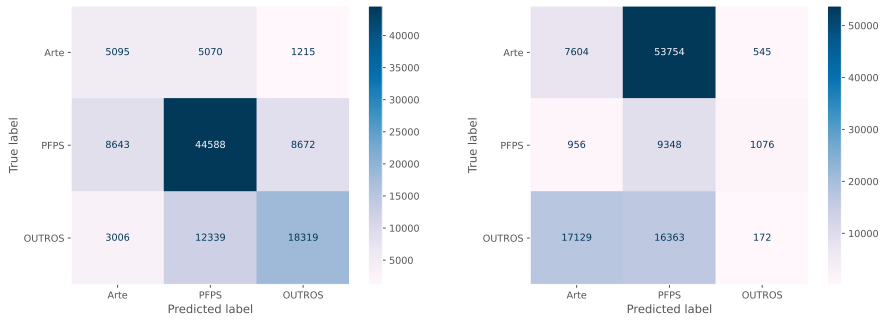


Figura 1. Valor Médio e Intervalo de Confiança

o intervalo de confiança para os seis atributos mais relevantes usados na classificação. Cada barra representa a média dos valores em uma categoria específica. Os intervalos de confiança (IC), geralmente calculados a 95%, acompanham as barras. Observamos que, para todos os atributos analisados, o valor médio das coleções na categoria PFPs é superior ao das outras duas categorias, indicando uma maior relevância desse tipo de coleção.

4. Resultados

Utilizamos um conjunto de classificadores para identificar de forma ternária as categorias de coleções, considerando as duas maiores em número de coleções (i.e., *art* e PFPs) e todas as demais categorias minoritárias em uma única classe, identificada como *outros*. Os modelos escolhidos foram: Floresta Aleatória (RF) e Máquinas de Vetor de



(a) RF com Oversampling **(b) SVM**

Figura 2. Matriz de confusão

Suporte (SVM)[Breiman 1996]. Implementamos o balanceamento de classes utilizando *undersampling*, que remove aleatoriamente amostras da classe majoritária para igualar à minoritária, e *oversampling*, que suplementa as classes minoritárias com instâncias de dados sintéticos que seguem a distribuição das instâncias reais.

Para avaliar os classificadores, dividimos a base de dados em 75% para treino e 25% para teste, treinando os classificadores com e sem balanceamento de classes. Na Tabela 3, apresentamos os melhores resultados dos classificadores com e sem tratamento de desbalanceamento de classes. Os resultados mostram que o modelo RF com *oversampling* teve um desempenho equilibrado em todas as métricas, destacando-se na classificação das categorias *art* e outros. Comparativamente, o modelo SVM teve desempenho inferior, especialmente na categoria PFPs. Isso sugere que a técnica de *oversampling* melhora significativamente a performance da Floresta Aleatória, tornando-a mais adequada para essa classificação ternária.

Tabela 3. Desempenho dos modelos avaliados com as métricas Precisão (P), Revocação (R), F1-score (F1), Acurácia (Acc) e F1-macro.

Modelos	Art			PFPs			Outros			Acc	F1-macro
	P	R	F1	P	R	F1	P	R	F1		
KNN	0,70	0,68	0,69	0,38	0,20	0,26	0,51	0,61	0,56	0,61	0,50
SVM	0,68	0,87	0,76	0,60	0,09	0,16	0,67	0,51	0,58	0,67	0,50
RF (oversampling)	0,72	0,72	0,72	0,30	0,45	0,36	0,65	0,54	0,69	0,64	0,56

A Figura 2 reporta a matriz de confusão para os classificadores Random Forest que obteve o melhor desempenho e a matriz de confusão para o classificador SVM, que obteve o segundo melhor desempenho. Cada linha representa as coleções em uma classe real, enquanto cada coluna representa as coleções em uma classe prevista. Na Figura 2a, observa-se que o modelo RF enfrenta maior dificuldade na diferenciação entre as classes *Pfps* e *Outros*, bem como entre *Outros* e *Pfps*. Essas duas classes apresentaram o maior número de erros de predição para o modelo RF, indicando uma tendência de confusão entre elas. Já na Figura 2b observa-se que para o modelo SVM o maior desafio está na distinção entre as classes *Outros* e *Art*, e vice-versa. Essas categorias destacam-se como as mais problemáticas para o modelo SVM, evidenciando dificuldades específicas na classificação desses dados.

5. Considerações Finais

Este trabalho apresentou uma metodologia eficaz para a identificação e classificação de coleções de NFTs com base em características estruturais. A coleta e o pré-processamento de uma vasta quantidade de dados da *OpenSea* permitiram a seleção de coleções com atributos relevantes, essenciais para uma análise precisa. Nossos resultados destacam a importância de técnicas de balanceamento de classes e do uso de classificadores robustos para melhorar a performance na categorização das coleções.

Para o futuro, a construção de um classificador automático mais avançado promete não apenas aprimorar a acurácia e a confiabilidade das classificações, mas também expandir as possibilidades de análise no ecossistema de NFTs. Focar em temas como a identificação de estilos de autores e a detecção de cópias de NFTs enfatiza a relevância dessa pesquisa, contribuindo significativamente para a integridade e a organização do mercado de NFTs. A capacidade de categorizar automaticamente coleções de NFTs é crucial para otimizar a curadoria, melhorar a experiência do usuário e manter a confiança no mercado digital.

Referências

Aspembitova, A. T., Feng, L., and Chew, L. Y. (2021). Behavioral structure of users in cryptocurrency market. *PLOS ONE*, 16(1):1–19.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Casale-Brunet, S., Ribeca, P., Doyle, P., and Mattavelli, M. (2021). Networks of ethereum non-fungible tokens: A graph-based analysis of the erc-721 ecosystem. In *2021 IEEE International Conference on Blockchain*, pages 188–195. IEEE.

Costa, D., La Cava, L., and Tagarelli, A. (2023). Show me your nft and i tell you how it will perform: Multimodal representation learning for nft selling price prediction. In *Proceedings of the ACM Web Conference 2023*, pages 1875–1885.

Hu, T., Liu, X., Chen, T., Zhang, X., Huang, X., Niu, W., Lu, J., Zhou, K., and Liu, Y. (2021). Transaction-based classification and detection approach for ethereum smart contract. *Information Processing Management*, 58(2):102462.

Nadini, M., Alessandretti, L., Di Giacinto, F., Martino, M., Aiello, L. M., and Baronchelli, A. (2021). Mapping the nft revolution: market trends, trade networks, and visual features. *Scientific reports*, 11(1):20902.

Valadares, J. A., Villela, S. M., Bernardino, H. S., Gonçalves, G. D., and Vieira, A. B. (2023). Mapping user behaviors to identify professional accounts in ethereum using semi-supervised learning. *Expert Systems with Applications*, 229:120438.

Wang, Q., Li, R., Wang, Q., and Chen, S. (2021). Non-fungible token (nft): Overview, evaluation, opportunities and challenges.

Xu, G., Guo, B., Su, C., Zheng, X., and Liang, K. (2020). Am i eclipsed? a smart detector of eclipse attacks for ethereum. *Computers Security*, 88:101604.