DeFi Protocol Risk Analysis: An Approach Based on On-Chain Metrics and Classifier Committees

Edenilson do Nascimento Silva[1], **Isaac Lima Oliveira Filho**[2], **Amanda Gondim de Oliveira**[1]

[1]Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró – RN – Brasil

[2]Universidade do Estado do Rio Grande do Norte (UERN)
Mossoró – RN – Brasil

edenilson.silva19222@alunos.ufersa.edu.br

isaacoliveira@uern.br, amandagondim@ufersa.edu.br

Abstract. *DeFi protocols manage billions of dollars in digital assets while remaining exposed to risks such as liquidity collapse, smart contract exploitation, and rug pulls. This paper investigates on-chain financial and categorical metrics, aggregated at the protocol level, to classify DeFi protocol risk through Machine Learning. Using data from the DeFiLlama API, we compared individual classifiers, homogeneous and heterogeneous committees, a stacking meta-classifier, and K-Means. The best result, a heterogeneous committee combining Decision Tree and k-NN, achieved F1-Score of 0.4482 (Friedman test,* $p < 0.05$*), with algorithmic bias identified across blockchains.*

Resumo. *Protocolos DeFi movimentam bilhões de dólares em ativos digitais e permanecem expostos a riscos como colapso de liquidez, exploração de contratos inteligentes e rug pulls. Este artigo investiga métricas financeiras e categóricas on-chain, agregadas ao nível do protocolo, para classificar o risco de protocolos DeFi por meio de Aprendizagem de Máquina. Utilizando dados da API do DeFiLlama, comparamos classificadores individuais, comitês homogêneos e heterogêneos, um meta-classificador via stacking e K-Means. O melhor resultado, um comitê heterogêneo combinando Árvore de Decisão e k-NN, obteve F1-Score de 0,4482 (teste de Friedman,* $p < 0{,}05$*), com viés algorítmico identificado entre blockchains.*

1. Introdução

As Finanças Descentralizadas (Decentralized Finance – DeFi) representam um dos principais casos de uso da tecnologia blockchain, movimentando bilhões de dólares em ativos digitais por meio de serviços como empréstimos descentralizados, exchanges descentralizadas (DEXs), *yield farming* e *staking*. Apesar do crescimento acelerado do setor, diversos protocolos apresentam falências, perda de liquidez e golpes conhecidos como *rug pulls*, nos quais desenvolvedores drenam a liquidez de um protocolo, deixando investidores com tokens sem valor.

A transparência inerente às blockchains públicas cria oportunidades para o uso de Aprendizagem de Máquina na identificação de protocolos de risco. Contudo, a maior

parte da literatura opera em granularidade elevada – análise estática de *bytecode* de contratos inteligentes, grafos de transação em alta resolução temporal ou sequências de eventos dos primeiros instantes após o lançamento de um token –, exigindo infraestrutura de coleta sofisticada e dados raramente disponíveis de forma agregada. Há, portanto, espaço relevante para investigar até que ponto métricas já consolidadas ao nível do protocolo – como valor total bloqueado, capitalização de mercado, variações percentuais de liquidez e blockchain de operação, publicamente disponíveis via agregadores como o DeFiLlama – sustentam uma classificação de risco informativa sem necessidade de processar dados brutos de transação ou *bytecode*.

O objetivo deste artigo é analisar o risco de protocolos DeFi a partir dessas métricas agregadas, empregando Aprendizagem de Máquina supervisionada e não supervisionada como instrumento de investigação – e não como fim em si mesmo. Especificamente, busca-se: (i) identificar quais métricas agregadas estão mais associadas a comportamentos de risco; (ii) comparar estratégias de classificação – individual, comitês homogêneos, comitês heterogêneos e *stacking* – para determinar qual melhor extrai sinal de risco desse tipo de dado; (iii) validar estatisticamente as diferenças de desempenho observadas; (iv) avaliar a robustez e os vieses algorítmicos da melhor estratégia; e (v) investigar, via aprendizado não supervisionado, se a separação entre protocolos de alto e baixo risco decorre de uma estrutura natural nos dados. O artigo está organizado da seguinte forma: a Seção 2 discute os trabalhos relacionados; a Seção 3 descreve a metodologia; a Seção 4 apresenta e discute os resultados experimentais; e a Seção 5 conclui o trabalho.

2. Trabalhos Relacionados

A literatura sobre segurança e análise de risco em DeFi pode ser organizada por níveis de granularidade dos dados, o que ajuda a posicionar a contribuição deste trabalho. Em nível de código de contrato, [Kong et al. 2023] propõem o DeFiTainter para detectar manipulação de preço (precisão de 96% e *recall* de 91,3% em 1.195 protocolos), e [Wen et al. 2024] apresentam o FORAY, que abstrai contratos em Grafos de Fluxo de Tokens e sintetiza ataques em 27 de 34 incidentes reais. Ambos são eficazes, mas dependem de análise ao *bytecode* – granularidade muito mais profunda que a explorada aqui.

Em nível comportamental e transacional, [Ren et al. 2025] propõem o LookAhead, que identifica contratos adversários combinando *Transformers*, XGBoost, Random Forest e *stacking* (F1 = 0,8966) – evidência de que árvores combinadas via *stacking* lidam bem com o desbalanceamento, achado que nossos experimentos (Seção 4) parcialmente contestam, já que a votação majoritária simples superou o *stacking*. [Wu et al. 2024] propõem o TokenScout, com redes neurais de grafos temporais sobre as primeiras 500 transações (acurácia balanceada de 98,41%), exigindo mapeamento detalhado da rede de transferências – dado indisponível em agregadores como o DeFiLlama.

No nível de dados estruturados, mais próximo da abordagem aqui adotada, [Nguyen et al. 2023] constroem o RugPullBank (F1 = 99% com XGBoost sobre 23 características de pools da Uniswap), indicando que a especificidade das *features* impacta o poder preditivo. O SolRPDS [Alhaidari et al. 2025] evidencia que a arquitetura econômica da blockchain influencia o comportamento dos golpistas, motivando a inclusão da variável *chain*. Abordagens temporais incluem [Behfar et al. 2026], que combina grafos convolu-

cionais e caminhadas aleatórias temporais na Ethereum, e [Green et al. 2024], que aplica Análise de Sobrevivência para modelar o tempo até a liquidação – direção apontada como trabalho futuro. [Tran et al. 2026] alerta para a fragilidade temporal (*concept drift*) de limiares fixos frente à evolução das táticas de ataque.

A revisão sistemática de [Luo et al. 2024] organiza a literatura pelo ciclo de vida do protocolo, destacando desbalanceamento extremo, escassez de rótulos e rápida adaptação dos fraudadores – desafios manifestados concretamente em nossos experimentos. Em síntese, a literatura concentra-se em fontes de alta granularidade, alcançando métricas superiores; a contribuição deste artigo não é superá-las, mas caracterizar empiricamente os limites de modelos construídos exclusivamente sobre métricas agregadas e publicamente acessíveis, incluindo análises de robustez e viés algorítmico ausentes na maior parte dos correlatos.

3. Metodologia

Esta seção descreve o pipeline experimental empregado para avaliar métricas agregadas na identificação de protocolos DeFi de alto risco. Após a coleta e o pré-processamento dos dados da API do DeFiLlama, o desbalanceamento foi tratado com SMOTE e os conjuntos divididos de forma estratificada. A modelagem seguiu de forma incremental – classificadores individuais de diferentes fundamentos matemáticos, comitês homogêneos (Bagging) e heterogêneos (Voting), meta-classificação via Stacking e agrupamento não supervisionado com K-Means (Figura 1) –, com as diferenças de desempenho submetidas aos testes de Friedman e Nemenyi.

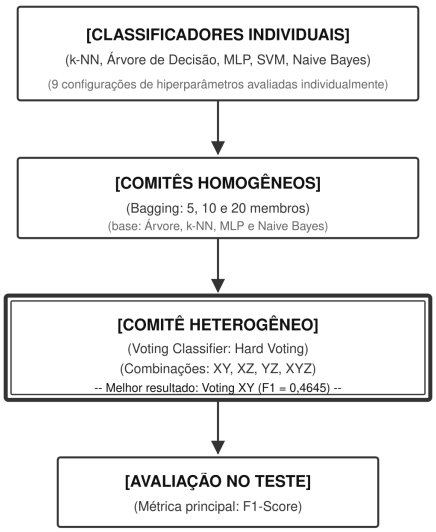


Figura 1. Fluxograma das estratégias de classificação avaliadas, dos classificadores individuais ao comitê heterogêneo.

3.1. Base de Dados

A base de dados foi construída a partir da API pública do DeFiLlama, o principal agregador de dados on-chain do ecossistema DeFi, contemplando um total de 7.631 protocolos. A variável alvo (`is_high_risk`) foi construída de forma binária, assumindo valor 1 para

protocolos que sofreram eventos de *rug pull*, foram declarados inativos (`deadFrom`) ou apresentaram alertas de segurança (`warningBanners`), e valor 0 caso contrário. A base resultante apresenta um forte desbalanceamento entre classes: 7.018 protocolos (92,0%) de Baixo Risco e 613 protocolos (8,0%) de Alto Risco – proporção consistente com o desafio de escassez de exemplos de risco apontado por [Luo et al. 2024].

Os atributos preditores incluem métricas financeiras contínuas – valor total bloqueado (*tvl*), variação percentual em 1 hora, 1 dia e 7 dias, capitalização de mercado (*mcap*), capital em *staking* e em *pool2* – e a blockchain de operação do protocolo (*chain*), tratada como variável categórica.

As métricas preditoras foram coletadas em um único *snapshot* contemporâneo à extração dos dados, e não em instante anterior ao evento de risco. Para protocolos já inativos ou vítimas de *rug pull*, portanto, os valores refletem parcialmente o estado do protocolo após o evento – implicação metodológica discutida na Seção 4.9.

3.2. Pré-processamento

O pré-processamento incluiu as seguintes etapas: (i) imputação por constante (zero) para valores ausentes nas métricas financeiras, justificada pelo fato de que, no domínio DeFi, a ausência de registro indica a inexistência do serviço, e não um erro de coleta; (ii) tratamento de *outliers* via Winsorização, com *clipping* nos percentis 1% e 99%, mitigando a distorção causada por protocolos com TVL na casa de bilhões de dólares; (iii) padronização Z-Score (`StandardScaler`) das variáveis numéricas contínuas, essencial para algoritmos sensíveis a escala como k-NN, SVM e MLP; e (iv) codificação *one-hot* da variável *chain*, considerando as cinco blockchains mais frequentes e agrupando as demais na categoria *Other*.

A partir desse pré-processamento, geraram-se quatro versões da base: a Original Tratada, com 13 atributos (7 financeiros e 6 *dummies* de rede), e três reduzidas via PCA (2, 3 e 4 componentes). Os experimentos reportados usam a base Original Tratada, por preservar a interpretabilidade direta das variáveis, essencial à análise de risco (Seção 4).

3.3. Divisão da Base e Balanceamento

A base foi dividida em 70% para treino e 30% para teste, com amostragem estratificada para preservar a proporção entre classes em ambos os conjuntos. Para mitigar o desbalanceamento, foi aplicada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) exclusivamente sobre o conjunto de treino, dentro de um *pipeline* que garante a ausência de vazamento de dados (*data leakage*) durante a validação cruzada. Todos os modelos foram avaliados via validação cruzada estratificada de 10 dobras (`StratifiedKFold`), complementada por uma avaliação final no conjunto de teste *hold-out*.

3.4. Estratégias de Classificação Avaliadas

A modelagem foi estruturada em cinco etapas incrementais, abrangendo abordagens supervisionadas e uma não supervisionada.

Classificadores Individuais. Foram avaliados cinco algoritmos de naturezas matemáticas distintas, com variações de hiperparâmetros totalizando nove configurações: (i) k-NN, com $k \in \{5, 11\}$ (proximidade geométrica, distância euclidiana); (ii) Árvore de Decisão, profundidade máxima em $\{5, 10\}$ (partição lógica, `criterion=gini`); (iii)

Naive Bayes gaussiano (probabilidades condicionais sob independência); (iv) Perceptron Multicamadas (MLP), arquiteturas $(100, 50)$ e $(200, 100)$, ativação `relu` e otimizador `adam`; e (v) SVM, com kernels linear e RBF ($C = 1{,}0$).

Comitês Homogêneos (Bagging). Com o objetivo de reduzir variância e mitigar overfitting, aplicou-se *Bootstrap Aggregating* (`BaggingClassifier`) para quatro dos cinco algoritmos base, em três tamanhos: 5, 10 e 20 membros, totalizando 12 configurações. O SVM foi deliberadamente excluído desta etapa em razão do custo computacional proibitivo do Bagging combinado a kernels não lineares sobre bases de porte médio.

Comitês Heterogêneos (Voting). Para explorar a complementaridade de hipóteses geradas por modelos de naturezas distintas, implementou-se votação majoritária simples (*Hard Voting*, `VotingClassifier`) combinando os três algoritmos com melhor desempenho nos comitês homogêneos: Árvore de Decisão (modelo X, F1 CV = 0,3941), k-NN (modelo Y, F1 CV = 0,3575) e MLP (modelo Z, F1 CV = 0,3437). Foram avaliadas as combinações parciais (XY, XZ, YZ) e a combinação tripla (XYZ).

Meta-Classificador (Stacking). Construiu-se uma arquitetura de *Stacking* (`StackingClassifier`) na qual as previsões do melhor comitê heterogêneo (XY) e da combinação completa dos três classificadores (XYZ) serviram como *meta-features* de entrada para um estimador de nível superior. Utilizou-se Regressão Logística como meta-classificador final, por sua robustez e eficiência como ponderador linear das probabilidades base.

Aprendizado Não Supervisionado (K-Means). O algoritmo K-Means ($k = 2$) foi aplicado sobre a base de dados com propósito investigativo: verificar se a distinção entre protocolos de Alto e Baixo Risco emerge organicamente da estrutura intrínseca dos dados financeiros agregados, operando de forma cega em relação aos rótulos reais de classe.

A métrica de avaliação prioritária foi o F1-Score em detrimento da Acurácia – escolha metodologicamente mandatória dado o desbalanceamento extremo da base: um classificador ingênuo que previsse todos os protocolos como Baixo Risco atingiria 92% de acurácia mascarando total incapacidade preditiva real.

4. Resultados e Discussão

Esta seção apresenta e discute os resultados das cinco etapas de modelagem, seguidos da validação estatística das diferenças observadas e das análises de erros por blockchain, robustez e agrupamento não supervisionado.

4.1. Classificadores Individuais

A Tabela 1 apresenta os resultados das nove configurações de classificadores individuais, ordenadas por F1 CV.

A Árvore de Decisão com profundidade máxima 10 obteve o melhor F1 individual (F1 CV = 0,3844; F1 Teste = 0,3912). Um padrão sistemático emerge da tabela: Árvore, MLP e SVM combinam Recall elevado (acima de 0,88) com Precisão reduzida (inferior a 0,25), consequência do SMOTE, que amplifica a sensibilidade à classe minoritária às custas da precisão; o k-NN é mais equilibrado, porém com F1 inferior ao das árvores. Do

Tabela 1. Resultados dos classificadores individuais (StratifiedKFold, $k = 10$, com SMOTE)

Modelo	F1 CV	±Std	Prec CV	Rec CV	AUC CV
Árvore (max=10)	**0,3844**	0,0253	0,2459	0,8832	0,8509
Árvore (max=5)	0,3820	0,0279	0,2421	0,9065	0,8438
k-NN ($k = 11$)	0,3617	0,0142	0,2835	0,5035	0,7647
k-NN ($k = 5$)	0,3524	0,0235	0,2855	0,4638	0,7605
MLP (100,50)	0,3444	0,0162	0,2122	0,9159	0,8447
MLP (200,100)	0,3434	0,0193	0,2111	0,9229	0,8338
SVM (RBF)	0,2993	0,0163	0,1782	0,9369	0,7907
SVM (Linear)	0,2681	0,0147	0,1566	0,9323	0,7595
Naive Bayes	0,2657	0,0134	0,1551	0,9253	0,8135

ponto de vista aplicado, o Recall alto é desejável, pois não detectar um protocolo de Alto Risco tende a custar mais do que investigar um seguro desnecessariamente.

4.2. Comitês Homogêneos

A Tabela 2 apresenta os resultados dos 12 comitês homogêneos construídos via Bagging para quatro famílias de classificadores em três tamanhos.

Tabela 2. Resultados dos comitês homogêneos (BaggingClassifier)

Classificador	Membros	F1 CV	±Std	AUC CV	F1 Teste
Árvore	20	**0,3941**	0,0290	0,8725	0,4043
Árvore	5	0,3937	0,0293	0,8662	0,4005
Árvore	10	0,3926	0,0280	0,8698	0,4057
k-NN	20	0,3575	0,0245	0,8621	0,3714
k-NN	5	0,3543	0,0257	0,8558	0,3740
k-NN	10	0,3541	0,0256	0,8610	0,3728
MLP	5	0,3437	0,0165	0,8448	0,3511
MLP	20	0,3435	0,0162	0,8450	0,3494
MLP	10	0,3432	0,0174	0,8446	0,3511
Naive Bayes	10	0,2667	0,0136	0,8131	0,2641
Naive Bayes	20	0,2667	0,0132	0,8126	0,2633
Naive Bayes	5	0,2663	0,0130	0,8129	0,2635

O Bagging de Árvore com 20 membros obteve o melhor resultado (F1 CV = 0,3941), ganho de apenas 2,5% sobre o melhor individual. O tamanho do comitê teve impacto negligenciável dentro de cada família (a Árvore variou entre 0,3926 e 0,3941), confirmando que o benefício do Bagging é a redução de variância – não o aumento do poder preditivo médio –, ganho que satura quando os estimadores base já operam sobre amostras balanceadas por SMOTE. O MLP foi o mais estável entre dobras (F1 Std = 0,0162 com 20 membros), embora inferior à Árvore em desempenho médio.

Com base nesses resultados, os classificadores para a etapa heterogênea foram definidos como: **X** = Árvore de Decisão (Bagging 20 membros, F1 CV = 0,3941), **Y** = k-NN (Bagging 20 membros, F1 CV = 0,3575) e **Z** = MLP (Bagging 5 membros, F1 CV = 0,3437).

4.3. Comitês Heterogêneos

A Tabela 3 apresenta os resultados das quatro combinações heterogêneas avaliadas via *Hard Voting*.

O comitê XY (Árvore de Decisão + k-NN) obteve o melhor desempenho entre todas as estratégias avaliadas, com F1 CV = 0,4482 e F1 Teste = 0,4645 – ganho relativo de 13,7% em relação ao melhor comitê homogêneo e de 16,6% em relação ao melhor

Tabela 3. Resultados dos comitês heterogêneos (VotingClassifier, hard)

Comitê	F1 CV	±Std	Acc CV	F1 Teste
XY (Árvore + k-NN)	**0,4482**	0,0215	0,9073	**0,4645**
XZ (Árvore + MLP)	0,4018	0,0248	0,7909	0,4169
YZ (k-NN + MLP)	0,3931	0,0197	0,8770	0,4042
XYZ (Árvore + k-NN + MLP)	0,3765	0,0175	0,7579	0,3842

classificador individual. Notavelmente, a adição do MLP ao comitê (combinações XZ, YZ e XYZ) sistematicamente reduziu o desempenho. O ganho do XY decorre de um princípio central de comitês por votação: a combinação só supera seus membros quando os erros destes são descorrelacionados. A Árvore de Decisão (fronteiras de partição paralelas aos eixos) e o k-NN (decisão por proximidade local) possuem vieses indutivos distintos e, portanto, tendem a errar em instâncias diferentes, permitindo que a votação corrija equívocos isolados. O MLP, por outro lado, tende a errar nas mesmas regiões que o k-NN; na votação de três membros, esses dois votos correlacionados sobrepujam o voto correto da Árvore em parte das amostras – o que explica por que a combinação mais rica (XYZ) é, paradoxalmente, a de pior desempenho (Figura 2).

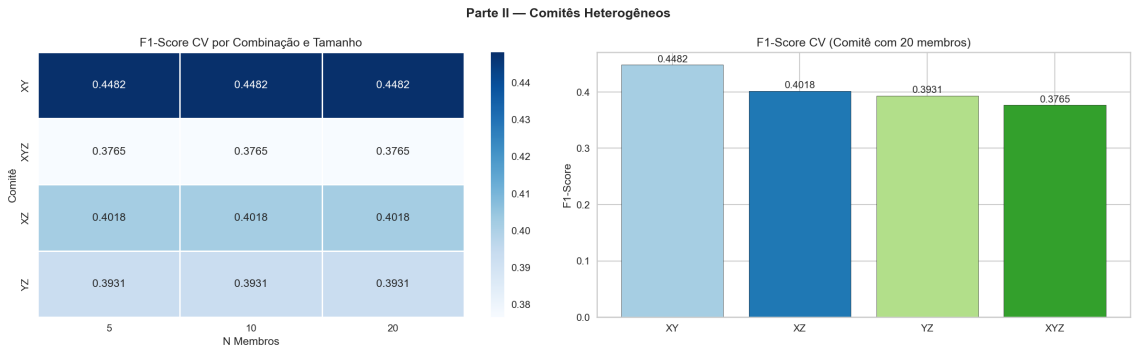


Figura 2. F1-Score dos comitês heterogêneos por combinação (XY, XZ, YZ e XYZ).

4.4. Meta-Classificador: Stacking

A Tabela 4 apresenta os resultados das duas configurações de *Stacking* avaliadas, com Regressão Logística como meta-classificador.

Tabela 4. Resultados do meta-classificador Stacking (meta-estimador: Regressão Logística)

Configuração	F1 CV	±Std	AUC CV	F1 Teste
Stacking-XY	0,3869	0,0259	0,8743	0,3953
Stacking-XYZ	0,3984	0,0271	0,8846	0,4073

O *Stacking* não superou o comitê heterogêneo XY por votação majoritária, apesar de ser teoricamente mais sofisticado. A Regressão Logística como meta-classificador obteve F1 CV de 0,3984 (Stacking-XYZ), contra 0,4482 do Voting XY – resultado contraintuitivo que diverge do observado por [Ren et al. 2025]. Uma possível explicação é que a Regressão Logística, ao aprender uma combinação linear das probabilidades base, penaliza a complementaridade discreta entre Árvore e k-NN de forma ineficiente; a votação majoritária, por outro lado, preserva diretamente essa diversidade. A validação estatística na Seção 4.5 nuança essa conclusão.

4.5. Validação Estatística

Para verificar se as diferenças de desempenho entre as estratégias são estatisticamente significativas, aplicou-se o teste de Friedman sobre os F1-Scores obtidos em cada uma das 10 dobras da validação cruzada. O teste rejeitou a hipótese nula de equivalência entre os modelos ($\chi^2 = 23{,}00$; $p = 0{,}00004$), indicando diferença estatisticamente significativa entre, pelo menos, um par de estratégias.

O pós-teste de Nemenyi, aplicado para identificar quais pares diferem entre si, revelou que o comitê heterogêneo XY é estatisticamente superior tanto ao algoritmo individual ($p < 0{,}001$) quanto ao comitê homogêneo ($p = 0{,}022$). Entretanto, a diferença entre XY e o *stacking* não se mostrou estatisticamente significativa ($p = 0{,}402$), apesar da diferença numérica de F1-Score. Esse resultado adiciona uma nuance importante: embora o Voting XY tenha apresentado o maior F1-Score médio, não há evidência estatística suficiente, neste conjunto de dados, para afirmar que ele é genuinamente superior ao *stacking* – apenas que ambos são superiores ao algoritmo individual isolado.

4.6. Análise de Erros e Viés Algorítmico

A matriz de confusão do comitê heterogêneo XY no conjunto de teste revelou 89 Falsos Negativos e 130 Falsos Positivos, de um total de 2.290 protocolos. Em termos operacionais, esses valores correspondem a uma Precisão de aproximadamente 0,42 para a classe de Alto Risco: dos 225 protocolos sinalizados, 130 (58%) são falsos alarmes. Tal taxa limita o uso do comitê a uma triagem inicial de baixo custo, na qual a inspeção manual de um protocolo seguro é tolerável, e não a uma decisão automatizada de exclusão – compromisso coerente com a priorização deliberada do Recall, já que, em triagem de risco, deixar de sinalizar um protocolo efetivamente arriscado é mais custoso do que investigar um falso positivo. A Tabela 5 apresenta a taxa de Falso Negativo desagregada por blockchain de operação, revelando viés algorítmico expressivo.

Tabela 5. Taxa de Falso Negativo por blockchain (comitê heterogêneo XY)

Blockchain	Total Alto Risco	Falsos Negativos	Taxa FN
Binance	11	8	72,7%
Other	108	55	50,9%
Ethereum	18	9	50,0%
Multi-Chain	28	13	46,4%
Solana	11	3	27,3%
Arbitrum	8	1	12,5%

Enquanto protocolos na rede Arbitrum apresentaram taxa de Falso Negativo de apenas 12,5%, protocolos na Binance atingiram 72,7%. Esse viés é, em parte, explicado pelo tamanho reduzido da amostra de Alto Risco em algumas redes (apenas 11 casos em Binance e Solana), limitando a capacidade do modelo de generalizar padrões específicos durante o treinamento, mesmo após o balanceamento via SMOTE. Esse achado é relevante: modelos de risco aplicados uniformemente entre blockchains podem oferecer garantias de detecção substancialmente diferentes a depender da rede – aspecto raramente discutido nos trabalhos correlatos, que em sua maioria concentram-se em uma única blockchain (tipicamente Ethereum).

4.7. Análise de Robustez

Para avaliar a sensibilidade do comitê XY a perturbações nos dados, foram conduzidos três experimentos complementares: adição de ruído gaussiano, remoção de atributos (*fe-*

ature ablation) e reamostragem em subconjuntos.

A adição de ruído gaussiano às variáveis numéricas, mesmo em magnitude pequena (5% do desvio padrão de cada *feature*), provocou queda abrupta de F1 = 0,4645 para F1 = 0,0548 – degradação de 88%. O ruído foi aplicado exclusivamente às sete variáveis numéricas do conjunto de teste – já Winsorizadas e padronizadas (Z-Score) na etapa de pré-processamento –, mantendo o modelo treinado sobre dados limpos; para cada variável, amostrou-se ruído gaussiano de média zero e desvio igual a 5% do desvio padrão da própria variável. Um diagnóstico controlado afastou a hipótese de artefato metodológico: a padronização foi confirmada (média ≈ 0, desvio ≈ 1) e a degradação persistiu ao isolar cada classificador. A investigação isolada de cada componente revelou que a Árvore de Decisão é a principal responsável (queda de 84% no F1 isolado, contra 34% do k-NN). Cortes de decisão rígidos tornam a árvore vulnerável a deslocamentos mínimos de instâncias entre regiões de decisão; o k-NN, por basear suas decisões em proximidade métrica, demonstrou maior tolerância a pequenas perturbações. Esse resultado expõe um *trade-off* relevante: o melhor comitê em dados limpos é também o mais frágil a imprecisões de entrada – cenário plausível em aplicações reais com dados de API.

O experimento de *feature ablation* funciona também como uma análise de importância dos atributos (explicabilidade): indicou que o *tvl* é a variável mais determinante – sua remoção reduziu o F1 de 0,4645 para 0,3896 (queda de 16,1%). Esse resultado é coerente com o domínio, uma vez que o valor total bloqueado é o principal indicador da saúde econômica de um protocolo, sendo seu colapso o sinal de risco mais forte disponível entre as métricas agregadas. A remoção de *change_7d* e *staking* produziu quedas modestas (≈3%), enquanto a remoção de *mcap* resultou em leve melhora (F1 = 0,4850), sugerindo redundância informacional com as demais métricas financeiras. A reamostragem em 10 subconjuntos estratificados produziu F1 médio de 0,4799 com desvio padrão de 0,0251 (CV = 5,2%), indicando estabilidade satisfatória na ausência de ruído artificial. Os três experimentos estão sintetizados na Figura 3, evidenciando que a fragilidade do modelo é específica a perturbações no valor das *features*, não à composição da amostra de treino.

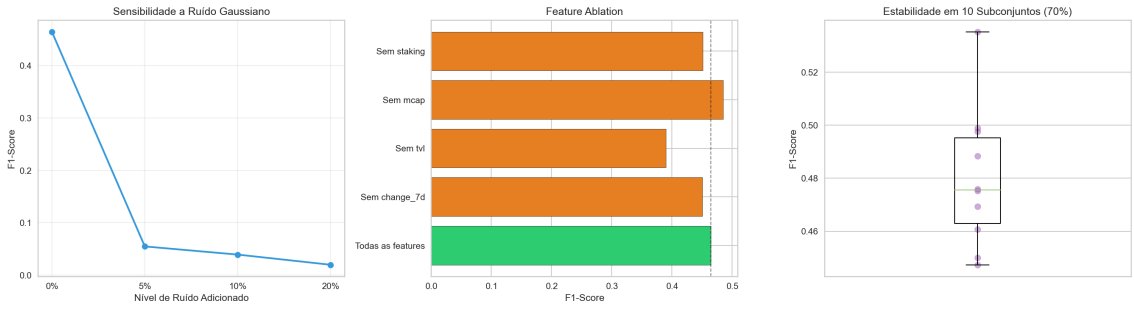


Figura 3. Painel de análise de robustez do comitê XY: (a) degradação do F1-Score sob ruído gaussiano; (b) impacto da remoção individual de atributos (*feature ablation*); (c) distribuição do F1-Score na reamostragem em subconjuntos estratificados.

4.8. Algoritmo Não Supervisionado

A aplicação do K-Means cumpre um papel investigativo específico: verificar se a dificuldade de classificação observada nas seções anteriores decorre de uma limitação dos

modelos supervisionados ou da ausência de estrutura separável nos próprios dados. Funciona, portanto, como um experimento de controle para a hipótese de que o risco emergiria naturalmente do agrupamento das métricas agregadas, sem necessidade de rótulos.

O algoritmo K-Means, aplicado com $k = 2$ sobre a base Original Tratada, obteve *Adjusted Rand Index* de $-0{,}0018$, *Normalized Mutual Information* de $0{,}0003$ e *Silhouette Score* de $0{,}2991$. O ARI próximo de zero indica que os *clusters* formados não correspondem à divisão real entre classes de risco, resultado coerente com a projeção PCA-2D na qual as classes aparecem visualmente sobrepostas (Figura 4). Conclui-se que o risco de um protocolo DeFi não decorre de uma estrutura de agrupamento intrínseca nas métricas agregadas, o que não apenas reforça a necessidade do aprendizado supervisionado, como também ajuda a explicar o teto de desempenho modesto obtido pelos classificadores: o sinal discriminativo é genuinamente fraco e entrelaçado entre as classes, e não meramente subaproveitado pelos modelos – constatação que dialoga com a observação de [Tran et al. 2026] sobre a sofisticação crescente das táticas de evasão, que tendem a camuflar protocolos fraudulentos entre os legítimos.

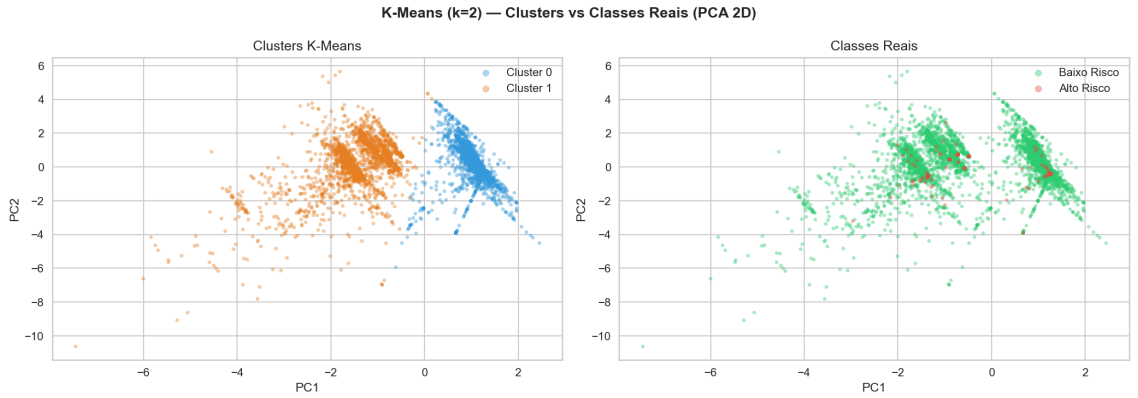


Figura 4. Comparação entre os *clusters* encontrados pelo K-Means ($k = 2$) e as classes reais de risco, projetados no espaço PCA-2D.

4.9. Discussão à Luz da Literatura

Comparando os resultados obtidos (F1 Teste = 0,4645) com os trabalhos da Seção 2, observa-se diferença substancial em relação a abordagens de maior granularidade: o RugPullBank de [Nguyen et al. 2023] reporta F1 de 99% com 23 características específicas de pools da Uniswap; o TokenScout de [Wu et al. 2024] alcança acurácia balanceada de 98,41%; e o LookAhead de [Ren et al. 2025] obtém F1 de 0,8966 com análise de *bytecode*. Essa diferença deve ser interpretada como evidência de um *trade-off* entre granularidade dos dados e poder preditivo: métricas agregadas, embora publicamente acessíveis e computacionalmente baratas, carregam consideravelmente menos sinal discriminativo do que características extraídas de código ou grafos de transação. Ainda assim, o ganho estatisticamente significativo do comitê XY em relação às abordagens individuais e homogêneas, combinado com as análises de viés e robustez – ausentes na maioria dos trabalhos correlatos –, posiciona este estudo como uma caracterização empírica relevante dos limites e do comportamento de modelos de risco DeFi construídos sobre dados de acesso público, cenário realista para investidores e ferramentas de triagem sem infraestrutura própria de indexação on-chain.

Os três desafios centrais destacados pela revisão de [Luo et al. 2024] – e antecipados ao final da Seção 2 – manifestaram-se concretamente neste estudo: o desbalanceamento extremo de classes exigiu reamostragem via SMOTE e limitou a Precisão alcançável; a escassez de rótulos confiáveis reflete-se na definição heurística do alvo a partir de sinais indiretos (*rug pull*, inatividade e alertas de segurança); e a rápida adaptação dos fraudadores encontra eco tanto na fragilidade temporal quanto na sobreposição observada entre protocolos de risco e legítimos, que impõe um teto ao sinal extraível de métricas agregadas.

Uma consideração metodológica importante diz respeito ao alinhamento temporal entre os atributos e o rótulo. Como as métricas foram extraídas em um *snapshot* contemporâneo à observação, e não anterior ao evento de risco, para protocolos já falhos elas incorporam, em parte, uma assinatura do estado pós-evento: cerca de 93% dos protocolos de Alto Risco apresentam as variáveis de variação de preço (*change_1h*, *change_1d* e *change_7d*) no valor nulo imputado – indicativo de ausência de negociação –, contra aproximadamente 27% dos de Baixo Risco. Rigorosamente, portanto, o trabalho deve ser interpretado como uma *caracterização* do risco a partir da pegada on-chain atual do protocolo, e não como uma *previsão* anterior ao evento. Cabe ressaltar, contudo, que essa contaminação é limitada e não trivializa a tarefa: a mesma assinatura de inatividade ocorre em uma parcela relevante de protocolos legítimos de baixo volume, de modo que ela não separa as classes de forma limpa. Uma evidência direta disso é o próprio desempenho modesto obtido (F1 = 0,4645): caso o *snapshot* codificasse trivialmente o rótulo, o F1 se aproximaria dos valores próximos de 99% reportados pelas abordagens baseadas em *bytecode* e grafos de transação da Seção 2, o que não se observa. Adicionalmente, essa dependência do modelo em valores exatos das variáveis de variação ajuda a explicar a fragilidade a ruído documentada na Seção 4.7: pequenas perturbações deslocam essas variáveis de seu valor imputado, dissolvendo justamente o principal sinal discriminativo de que a Árvore de Decisão dispõe.

5. Conclusão

Este trabalho investigou a viabilidade de métricas on-chain agregadas ao nível do protocolo para a análise de risco em DeFi. A combinação heterogênea de Árvore de Decisão e k-NN por votação majoritária foi a estratégia estatisticamente mais eficaz, superando de forma significativa os algoritmos individuais e os comitês homogêneos (teste de Friedman, $p < 0{,}05$). A análise de robustez revelou fragilidade relevante: a dependência do componente baseado em árvore torna o modelo sensível a ruído na entrada – limitação prática para cenários de implantação real. A análise de erros expôs viés algorítmico entre blockchains, com Falso Negativo de 12,5% a 72,7%; e o K-Means confirmou que a separação entre classes não decorre de estrutura natural de agrupamento, validando a escolha pelo aprendizado supervisionado.

Como trabalhos futuros, pretende-se: (i) investigar a incorporação de variáveis temporais e estruturais [Luo et al. 2024, Wu et al. 2024] como atributos adicionais; (ii) explorar a modelagem do risco como função do tempo, nos moldes da Análise de Sobrevivência de [Green et al. 2024]; (iii) mitigar a fragilidade a ruído por meio da substituição da Árvore de Decisão por variantes mais robustas, como Gradient Boosting regularizado; e (iv) replicar os experimentos sobre as bases reduzidas via PCA, avaliando o impacto da redução de dimensionalidade sobre o desempenho e a robustez do comitê XY.

Referências

Alhaidari, A., Kalal, B., Palanisamy, B., and Sural, S. (2025). SolRPDS: A dataset for analyzing rug pulls in Solana decentralized finance. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, CODASPY '25, pages 293–298. ACM.

Behfar, S., Mortier, R., and Crowcroft, J. (2026). Probabilistic sampling-enhanced temporal-spatial GCN: A scalable framework for transaction anomaly detection in Ethereum networks. In *Proceedings of the Sixth European Workshop on Machine Learning and Systems*, EuroMLSys '26, pages 263–273. ACM.

Green, A. M., Giannattasio, M. P., Erickson, J. S., Seneviratne, O., and Bennett, K. P. (2024). DeFi survival analysis: Insights into the emerging decentralized financial ecosystem. *Distributed Ledger Technologies: Research and Practice*, 3(1):1–23.

Kong, Q., Chen, J., Wang, Y., Jiang, Z., and Zheng, Z. (2023). DeFiTainter: Detecting price manipulation vulnerabilities in DeFi protocols. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2023, pages 1144–1156, New York, NY, USA. ACM.

Luo, B., Zhang, Z., Wang, Q., Ke, A., Lu, S., and He, B. (2024). AI-powered fraud detection in decentralized finance: A project life cycle perspective. *ACM Computing Surveys*, 57(4):1–38.

Nguyen, M. H., Huynh, P. D., Dau, S. H., and Li, X. (2023). Rug-pull malicious token detection on blockchain using supervised learning with feature engineering. In *Proceedings of the 2023 Australasian Computer Science Week*, ACSW '23, pages 72–81. ACM.

Ren, S., He, L., Tu, T., Wu, D., Liu, J., Ren, K., and Chen, C. (2025). LookAhead: Preventing DeFi attacks via unveiling adversarial contracts. *Proceedings of the ACM on Software Engineering*, 2(FSE):1847–1869.

Tran, M. T., Killeen, B., and McGrath, T. (2026). Adaptive detection of DeFi SLID scams: A data-driven and industry-oriented framework for large-scale DeFi security. In *Proceedings of the 2026 Australasian Information Security Conference*, AISC '26. ACM.

Wen, H., Liu, H., Song, J., Chen, Y., Guo, W., and Feng, Y. (2024). FORAY: Towards effective attack synthesis against deep logical vulnerabilities in DeFi protocols. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pages 1144–1156, New York, NY, USA. ACM.

Wu, C., Chen, J., Zhao, Z., He, K., Xu, G., Wu, Y., Wang, H., Li, H., Liu, Y., and Xiang, Y. (2024). TokenScout: Early detection of Ethereum scam tokens via temporal graph learning. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pages 956–970, New York, NY, USA. ACM.