Análise estatística de base de dados de detecção de lavagem de dinheiro com uso de LLM

1st Thiago Cavalcanti Silva Barros
Programa de Pós-Graduação em Ciência da Computação
Universidade Federal do Agreste de Pernambuco
Garanhuns, Brasil
0009-0009-2397-1049

***Abstract*—Este trabalho apresenta uma análise estatística inferencial aplicada a dados de transações financeiras com foco na detecção de lavagem de dinheiro. Utilizando a base de dados AML disponibilizada pela IBM, foram investigadas diferenças entre transações legítimas e suspeitas por meio de estimadores pontuais, intervalos de confiança e testes de hipótese. Além disso, empregou-se um modelo de linguagem de grande porte (LLM), executado localmente, para atribuição heurística de rótulos de risco às transações, permitindo avaliar sua concordância com os rótulos reais da base. Os resultados indicam que transações suspeitas apresentam valores significativamente superiores aos das transações legítimas, com evidência estatística robusta mesmo sob distribuições altamente assimétricas. Observou-se forte correlação entre o valor da transação e o escore de risco atribuído pela LLM, refletindo o comportamento do modelo, enquanto não foi identificada correlação significativa entre valor e frequência de transações nesta base de dados. A análise de concordância revelou alinhamento fraco entre a classificação da LLM e os rótulos reais, evidenciando limitações do uso direto de modelos genéricos para tarefas de classificação sensíveis, sem treinamento prévio ou ajustes. O estudo destaca tanto o potencial quanto as restrições do uso de LLMs em cenários de análise financeira, reforçando a importância de abordagens estatísticas rigorosas na validação de padrões observados.**

***Index Terms*—Inferência estatística; Detecção de lavagem de dinheiro; Teste de hipóteses; Intervalo de confiança; Análise de correlação; Grande Modelo de Linguagem (LLM)**

I. Introdução

A lavagem de dinheiro constitui um dos principais desafios enfrentados por sistemas financeiros modernos, estando diretamente associada a atividades criminosas como corrupção, tráfico e financiamento ilícito [1]. Instituições financeiras são obrigadas a monitorar transações e reportar atividades suspeitas [2], frequentemente operando sobre volumes massivos de dados, altamente desbalanceados [3] e com rótulos incompletos ou imperfeitos.

Tradicionalmente, a detecção de lavagem de dinheiro é tratada como um problema de classificação supervisionada, utilizando regras heurísticas ou modelos estatísticos e de aprendizado de máquina treinados a partir de dados históricos. No entanto, essa abordagem apresenta limitações importantes: a dependência de rótulos confiáveis, a rigidez das regras manuais e a dificuldade de capturar padrões complexos e contextuais presentes nas transações financeiras [4].

Nesse contexto, modelos de linguagem de grande porte (Large Language Models – LLMs) surgem como uma alternativa promissora. Diferentemente de classificadores tradicionais, LLMs são treinados em grandes volumes de dados textuais e possuem capacidade de raciocínio contextual, permitindo analisar descrições estruturadas de transações e inferir níveis de risco sem necessidade de treinamento específico para a tarefa [5]. Essa característica torna tais modelos especialmente interessantes como ferramentas auxiliares em cenários onde rótulos são escassos, ruidosos ou altamente desbalanceados, como é o caso da detecção de lavagem de dinheiro.

Este trabalho propõe uma análise estatística baseada em dados sintéticos, com características realistas, de transações financeiras, utilizando uma LLM como instrumento de inferência de risco em regime zero-shot. Em vez de focar no desempenho preditivo do modelo, o objetivo central é investigar, por meio de técnicas estatísticas clássicas, se as classificações e escores de risco inferidos pela LLM apresentam associações significativas com características quantitativas das transações, como valores monetários e frequência de operações.

A contribuição principal deste estudo reside na integração entre inferência estatística e modelos de linguagem, avaliando o comportamento da LLM sob a ótica de comparação de grupos, intervalos de confiança, correlação estatística e concordância com rótulos de referência. Dessa forma, o trabalho busca responder se os padrões identificados pelo modelo podem ser considerados estatisticamente relevantes, bem como discutir as limitações e implicações do uso de LLMs como ferramentas auxiliares em sistemas de prevenção à lavagem de dinheiro.

II. Perguntas de pesquisa

Com o objetivo de investigar padrões estatisticamente relevantes em transações financeiras e avaliar o uso de modelos de linguagem como ferramentas auxiliares na detecção de lavagem de dinheiro, este trabalho foi guiado pelas seguintes perguntas de pesquisa.

A. Há evidência estatística (α = 0,05) de que transações rotuladas como lavagem apresentam maior valor médio de transação que transações legítimas?

Transações associadas à lavagem de dinheiro são frequentemente descritas na literatura como envolvendo

valores elevados ou padrões atípicos [6]. No entanto, essa suposição não é trivial, uma vez que operações ilícitas podem ser fragmentadas em múltiplas transações de menor valor com o objetivo de evitar mecanismos de detecção [7]. Assim, a comparação direta entre os valores médios dos dois grupos permite avaliar empiricamente se essa diferença se manifesta de forma estatisticamente significativa na base analisada.

B. O intervalo de confiança de 95% para o valor médio de transações suspeitas exclui valores críticos (ex: acima de 50.000 unidades monetárias)?

A simples estimativa pontual do valor médio pode ocultar incertezas relevantes. A análise por meio de intervalos de confiança permite avaliar não apenas a magnitude média das transações suspeitas, mas também a incerteza associada a essa estimativa. Verificar se o intervalo de confiança exclui valores considerados críticos fornece uma interpretação diretamente aplicável a cenários de tomada de decisão e gestão de risco.

O valor crítico não foi escolhido arbitrariamente, mas definido com base no comportamento empírico da população legítima, garantindo interpretação estatística e aderência ao contexto de risco financeiro.

C. Existe correlação estatisticamente significativa entre valor da transação e indicadores de risco / frequência de transações?

A presença de correlação entre o valor monetário das transações e indicadores de risco pode sugerir padrões comportamentais relevantes, como concentração de valores elevados em contas com maior atividade transacional. No entanto, correlação não implica causalidade, sendo necessário interpretar os resultados com cautela. Esta análise busca identificar associações estatísticas e discutir suas possíveis implicações, sem assumir relações causais diretas.

D. Existe concordância estatisticamente significativa entre a classificação de lavagem atribuída pela LLM e o rótulo real da base?

Embora modelos de linguagem não sejam treinados especificamente para a tarefa de detecção de lavagem de dinheiro, sua capacidade de inferência contextual pode levar à identificação de padrões compatíveis com rótulos de referência. Avaliar o grau de concordância entre as classificações da LLM e os rótulos disponíveis permite investigar se o modelo apresenta comportamento alinhado com o ground truth, bem como identificar vieses, limitações e potenciais usos como ferramenta auxiliar de triagem.

III. METODOLOGIA

Esta seção descreve a configuração do experimento, com os dados utilizados, o ambiente computacional e as ferramentas adotadas para a condução dos experimentos estatísticos e computacionais deste trabalho.

- **Base de dados**: Os experimentos foram conduzidos utilizando a base de dados sintética de detecção de lavagem de dinheiro disponibilizada pela IBM Research, amplamente empregada em estudos acadêmicos na área de Anti-Money Laundering (AML)[1]. Essa base foi projetada para simular, de forma realista, transações financeiras envolvendo tanto atividades legítimas quanto ilícitas, preservando propriedades estatísticas observadas em cenários reais, ao mesmo tempo em que evita o uso de dados sensíveis.

  A base completa possui aproximadamente 8GB e é composta por múltiplos arquivos representando diferentes aspectos do sistema financeiro simulado, incluindo contas, entidades e transações. Neste trabalho, foram utilizados exclusivamente dois arquivos:
  - LI-Small_Trans.csv, contendo registros de transações financeiras, com informações como valor, moeda, formato de pagamento, instituições envolvidas e rótulo indicativo de lavagem de dinheiro;
  - LI-Small_accounts.csv, contendo informações associadas às contas bancárias, como identificadores de bancos e entidades.

  A escolha por utilizar apenas esses dois arquivos deve-se ao foco do estudo em padrões estatísticos diretamente observáveis nas transações financeiras e na agregação por conta, sendo suficientes para responder às perguntas de pesquisa propostas. Além disso, essa escolha contribui para a viabilidade computacional do experimento, dado o grande volume de dados disponível na base completa.

- **Modelo de Linguagem e Ferramenta de Execução**: Para a inferência de risco associada às transações financeiras, foi utilizado um modelo de linguagem de grande porte pré-treinado, especificamente o Mistral-7B-Instruct, em sua versão quantizada mistral:7b-instruct-q4_K_M. Essa versão foi selecionada por representar um equilíbrio adequado entre capacidade de inferência e custo computacional, permitindo execução local sem necessidade de recursos de hardware especializados ou serviços pagos.
  A execução do modelo foi realizada por meio da ferramenta Ollama, que possibilita o uso local de LLMs de forma simplificada, sem dependência de infraestrutura em nuvem. A escolha do Ollama foi motivada por sua facilidade de integração com scripts em Python, baixo impacto no sistema após a execução e suporte nativo a modelos quantizados, características alinhadas às restrições computacionais do experimento.
  O modelo foi utilizado exclusivamente em regime de inferência zero-shot, sem qualquer tipo de treinamento adicional ou ajuste fino. Ou seja, os dados foram classifi-

[1] https://github.com/IBM/AML-Data

cados sem nunca terem sido vistos, sem exemplos e sem treinamento específico para esse caso. Além disso, foi personalizado para executar com ”role” (papel) de usuário e temperatura zero. Seu papel no experimento foi o de atuar como um instrumento de inferência, produzindo classificações e escores auxiliares de risco a partir das informações estruturadas das transações, posteriormente analisados por meio de técnicas estatísticas clássicas.
Nesse contexto, o uso da LLM não se caracteriza como aprendizado supervisionado ou treinamento do modelo, mas como um processo de inferência estatística baseado em um modelo previamente treinado. A opção por não realizar treinamento adicional decorre do fato de que o objetivo do estudo não é a otimização do desempenho preditivo, mas a análise de padrões e associações em uma base de dados real, empregando a LLM como um instrumento de apoio à extração e organização de variáveis auxiliares. Assim, o modelo é utilizado como ferramenta analítica, e não como um sistema treinado especificamente para o domínio dos dados em estudo.

- **Ambiente computacional**: Todos os experimentos foram executados localmente em um computador pessoal com a seguinte configuração:
  - Memória RAM: 16GB;
  - Processador: : Intel Core i5-11400H (11ª geração), 2.70 GHz;
  - Placa de vídeo: NVIDIA GeForce GTX 1650. Essa configuração foi suficiente para executar o modelo selecionado de forma estável, embora com tempos de inferência relativamente elevados (entre 15 e 17 minutos, com recursos divididos entre a execução de outros programas, por exemplo navegador), o que influenciou decisões relacionadas ao tamanho das amostras analisadas.

    Todo o código desenvolvido para o pré-processamento dos dados, chamadas ao modelo de linguagem e análises estatísticas foi implementado em Python, utilizando Jupyter Notebook. O código completo do experimento encontra-se disponível para reprodutibilidade em um repositório do GitHub[2].

IV. Experimento

Esta seção descreve, de forma detalhada, as etapas realizadas no experimento, incluindo a definição das variáveis analisadas, o uso do modelo de linguagem para inferência de risco e as estratégias de amostragem adotadas para lidar com o forte desbalanceamento da base de dados.

A. Criação de variáveis estatisticamente interpretáveis

A partir dos dados de transações financeiras, foram construídas variáveis quantitativas com interpretação estatística direta, utilizadas nas análises subsequentes. A principal variável considerada foi o valor da transação, representado pelo montante monetário efetivamente transferido em cada operação. Como os dados envolvem múltiplas moedas e direções de pagamento, o valor absoluto da transação foi utilizado como medida unificada de magnitude financeira, sendo referido ao longo do trabalho como ”amount”.

Além disso, foi criada a variável frequência de transações (tx_frequency), definida como o número total de transações associadas a uma mesma conta de origem ao longo do período observado. Essa variável foi obtida por meio de agregação dos dados transacionais e tem como objetivo capturar padrões comportamentais relacionados à intensidade de uso das contas, frequentemente associados a atividades suspeitas em cenários de lavagem de dinheiro.

Essas duas variáveis (valor da transação e frequência de transações) constituem a base quantitativa para as análises estatísticas de comparação de grupos, estimação intervalar e correlação.

B. Rotulagem de risco com LLM

Para complementar os rótulos disponíveis na base de dados e explorar o uso de modelos de linguagem como instrumentos analíticos, foi empregada uma LLM pré-treinada de uso geral, não ajustada nem treinada especificamente para a tarefa de detecção de lavagem de dinheiro. Cada transação foi representada por um conjunto reduzido de atributos quantitativos, especificamente o valor da transação (amount) e a frequência de transações associadas à conta de origem (tx_frequency).

Essas informações foram incorporadas em um prompt textual padronizado, submetido ao modelo com o objetivo de obter uma avaliação qualitativa de risco. O prompt (figura 1) solicitava explicitamente a classificação binária da transação quanto à suspeita de lavagem de dinheiro (suspeita ou não suspeita), bem como a atribuição de um escore de risco discreto variando de 0 a 10, interpretado como uma medida auxiliar de propensão à lavagem de dinheiro.

```
Analyze the following transaction:

Amount: {row['amount']}
Transaction frequency (origin account): {row['tx_frequency']}

Tasks:
1. Is this transaction suspicious of money laundering? ('Yes' or 'No')
2. Assign a risk score from 0 to 10 (Integer)

Return ONLY valid JSON with:
is_laundering, risk_score
```

Fig. 1. Prompt utilizado.

O modelo retornou sua resposta exclusivamente no formato JSON, contendo os campos ”is_laundering” e ”risk_score”, que foram posteriormente incorporados ao conjunto de dados como variáveis auxiliares. Dado que o modelo foi utilizado exclusivamente em regime de inferência zero-shot, essas saídas foram tratadas como inferências independentes,

[2]https://github.com/ThiagoCavalcantiSilva/analise-estatistica-AML

utilizadas apenas para análise estatística e comparação com os rótulos de referência da base, sem qualquer retroalimentação ou ajuste do modelo.

C. Amostragem

Devido ao grande volume de dados disponível e às limitações computacionais associadas à execução local da LLM, as análises foram realizadas sobre amostras da base de transações. O tamanho da amostra foi fixado em 300 observações, valor considerado suficiente para permitir inferência estatística básica, possivelmente suficiente para detectar efeitos de magnitude moderada com nível de significância de 5%., ao mesmo tempo em que mantém tempos de execução viáveis para a etapa de inferência com o modelo de linguagem.

Além disso, a base de dados apresenta um forte desbalanceamento entre transações legítimas e transações rotuladas como lavagem de dinheiro, com a classe positiva representando uma fração muito pequena do total. Para lidar com esse aspecto, foram conduzidos dois experimentos complementares, com estratégias de amostragem distintas.

- **Experimento 1 - Amostragem aleatória simples**: No primeiro experimento, foi realizada uma amostragem aleatória simples de 300 transações, sem qualquer controle sobre a proporção das classes. Essa estratégia preserva, em média, a distribuição original da base de dados e permite observar o comportamento do modelo de linguagem e das variáveis estatísticas em um cenário que reflete mais fielmente a prevalência real de transações suspeitas.

- **Experimento 2 - Amostragem estratificada (80/20)**: No segundo experimento, foi adotada uma amostragem estratificada, garantindo que 20% da amostra fosse composta por transações rotuladas como lavagem de dinheiro e 80% por transações legítimas, totalizando novamente 300 observações (60 suspeitas e 240 legítimas). Essa estratégia não reflete a distribuição real da população, mas permite assegurar representatividade mínima da classe positiva, visando avaliar o quão assertiva é a classificação da LLM sem nenhum treinamento específico.
  A adoção da amostragem estratificada é fundamental para possibilitar comparações estatísticas entre grupos, cálculo de métricas de concordância e análise mais robusta do comportamento da LLM frente a transações efetivamente rotuladas como lavagem. Os resultados obtidos nesse cenário são interpretados como evidência em um contexto experimental controlado, e não como estimativas diretas da prevalência populacional.

V. Resultados e evidências

A. Comparação entre transações legítimas e de lavagem

Para investigar se transações rotuladas como lavagem de dinheiro apresentam valores significativamente maiores do que transações legítimas, as observações foram inicialmente separadas em dois grupos, com base no rótulo de lavagem presente na base de dados: (i) transações legítimas e (ii) transações associadas à lavagem de dinheiro.

Como etapa preliminar, avaliou-se a normalidade da variável amount em ambos os grupos. Valores financeiros costumam apresentar distribuições assimétricas, com caudas longas e presença de outliers, o que pode violar os pressupostos de testes paramétricos clássicos. Portanto, o teste de Shapiro–Wilk foi aplicado a amostras dos dois subconjuntos, considerando as limitações do próprio teste para grandes tamanhos amostrais.

A hipótese nula do teste estabelece que os dados seguem uma distribuição normal, enquanto a hipótese alternativa indica violação dessa suposição. Os resultados obtidos indicaram valores extremamente baixos da estatística W ($\approx 0,21$ para o grupo de lavagem e $\approx 0,12$ para o grupo legítimo), acompanhados de p-valores da ordem de 10^{-15}, muito inferiores ao nível de significância adotado ($\alpha = 0{,}05$). Dessa forma, a hipótese nula de normalidade foi rejeitada com evidência estatística extremamente forte, confirmando que os dados não seguem uma distribuição normal (um comportamento esperado em dados financeiros).

Diante da violação severa da suposição de normalidade, optou-se pelo uso de um teste não paramétrico para a comparação entre os grupos. Especificamente, foi aplicado o teste de Mann–Whitney U, adequado para comparar duas amostras independentes sem assumir normalidade, sendo robusto à presença de assimetria e outliers. O teste foi formulado de forma unilateral, conduzido considerando como hipótese nula que a distribuição dos valores das transações rotuladas como lavagem não é superior à das transações legítimas, e como hipótese alternativa que os valores das transações suspeitas tendem a ser maiores.

O teste foi formulado de forma unilateral, conduzido considerando como hipótese nula que a distribuição dos valores das transações rotuladas como lavagem não é superior à das transações legítimas, e como hipótese alternativa que os valores das transações suspeitas tendem a ser maiores. O resultado do teste apresentou uma estatística U elevada e um p-valor extremamente pequeno ($p \approx 2{,}44 \times 10^{-124}$), indicando que a probabilidade de se observar um resultado dessa magnitude sob a hipótese nula é praticamente nula.

Portanto, rejeita-se fortemente a hipótese nula ao nível de significância de 5%. Assim, há evidência estatística extremamente robusta de que transações rotuladas como lavagem de dinheiro apresentam valores significativamente maiores do que transações legítimas na base analisada. Esse resultado sugere que, na base analisada, transações suspeitas tendem a envolver montantes financeiros mais elevados, corroborando a existência de padrões monetários distintos entre os dois grupos.

A Figura 2 ilustra graficamente esse resultado, evidenciando que a distribuição dos valores das transações rotuladas como lavagem apresenta mediana e quartis superiores aos das transações legítimas, mesmo em escala

logarítmica. Observa-se também uma maior concentração de valores elevados no grupo de lavagem, corroborando o resultado do teste estatístico e a consequente rejeição da hipótese nula.

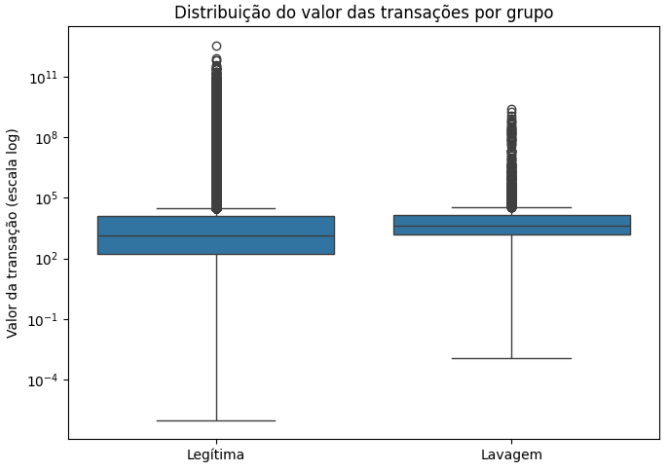

Fig. 2. Distribuição de valores das transações por grupo.

B. Intervalo de confiança de 95%

Esta questão investigou se o intervalo de confiança de 95% para o valor médio das transações rotuladas como suspeitas exclui valores considerados críticos do ponto de vista de risco financeiro. Dada a forte assimetria e a presença de outliers nos valores de transações, foi adotada uma abordagem não paramétrica baseada em bootstrap para a estimação do intervalo de confiança da média.

Diferentemente da utilização de limiares arbitrários, o valor crítico de referência foi definido de forma empírica, como o percentil 95 da distribuição dos valores das transações legítimas. Esse critério é amplamente empregado em contextos de detecção de anomalias e análise de risco, por representar o limite superior esperado para a grande maioria das transações normais.

O intervalo de confiança bootstrap de 95% obtido para o valor médio das transações suspeitas foi aproximadamente [1.842.071,81 ; 5.862.680,15] unidades monetárias, enquanto o valor crítico foi 600.226,34. Os resultados indicaram que o intervalo de confiança de 95% para o valor médio das transações suspeitas encontra-se inteiramente acima desse valor crítico, excluindo a faixa típica observada nas transações legítimas. Esse achado fornece evidência estatística forte de que transações classificadas como suspeitas apresentam valores médios substancialmente elevados, reforçando sua associação com maior risco de lavagem de dinheiro.

C. Correlação entre valor x risco e valor x frequência

a) Valor da transação × Escore de risco (LLM): O escore de risco é uma variável quantitativa ordinal no intervalo [0,10], gerada pela inferência da LLM, que representa o nível relativo de risco de lavagem de dinheiro atribuído a uma transação individual, com base em padrões linguísticos e heurísticos. Ou seja,a LLM não calcula risco estatístico, mas sim, interpreta o contexto da transação (valor, frequência, etc.), associa esses atributos a padrões típicos de detecção de lavagem de dinheiro e produz uma avaliação qualitativa, convertida em número.

Como o risco é ordinal e a distribuição é assimétrica, foi utilizado o coeficiente de correlação de Spearman, onde foi avaliada a existência de associação monotônica entre o valor da transação e o escore de risco atribuído pela LLM. Os resultados indicaram coeficientes elevados ($p \approx 0.79$ e $p \approx 0.82$), acompanhados de valores de p extremamente baixos $p < 10^{-60}$, indicando associação estatisticamente significativa em ambas as amostras e permitindo rejeitar a hipótese nula de ausência de correlação.

Esses resultados fornecem forte evidência estatística de uma associação monotônica positiva entre o valor da transação e o escore de risco atribuído pelo modelo, indicando que transações de maior valor tendem a receber classificações de risco mais elevadas. Esse padrão é consistente com a hipótese de que o valor da transação é um dos principais fatores utilizados pela LLM na avaliação de risco.

Entretanto, essa associação não deve ser interpretada como evidência de causalidade, uma vez que o escore de risco é uma variável derivada e o modelo foi explicitamente instruído com o valor da transação (o valor foi fornecido como entrada para a LLM), podendo introduzir viés na construção do julgamento. Assim, a correlação observada reflete o comportamento do modelo de linguagem, e não necessariamente os mecanismos reais subjacentes à lavagem de dinheiro. Além disso, a magnitude e estabilidade do coeficiente entre a amostra proporcional e a amostra estratificada sugerem que o resultado é robusto à estratégia de amostragem.

b) Valor da transação × Frequência de transações: A relação entre o valor da transação e a frequência de transações da conta de origem também foi avaliada utilizando o coeficiente de Spearman. No experimento com amostragem aleatória representativa da população real, obteve-se $p \approx 0.037$ ($p-valor \approx 0.521$). No experimento estratificado, os resultados foram semelhantes, com $p \approx -0.050$ ($p-valor \approx 0.389$).

O coeficiente de Spearman apresenta sinal negativo na amostra estratificada, indicando uma tendência monotônica inversa extremamente fraca entre valor da transação e frequência de transações. No entanto, dada a baixa magnitude do coeficiente, a ausência de significância estatística e a inconsistência do sinal entre amostras, esse resultado não fornece evidência de uma associação monotônica sistemática entre as variáveis.

Em ambos os casos, os valores de p não permitiram rejeitar a hipótese nula de ausência de correlação, indicando que não há evidência estatística de associação monotônica entre valor da transação e frequência operacional. Esse resultado sugere que contas que realizam transações com maior frequência não necessariamente movimentam valores mais elevados,

apontando para uma relativa independência entre intensidade operacional e volume financeiro no conjunto de dados analisado.

D. Concordância na classificação

A concordância entre a classificação binária produzida pela LLM e o rótulo real disponibilizado na base de dados da IBM foi avaliada por meio do coeficiente Kappa de Cohen, que mede o grau de acordo entre dois classificadores descontando a concordância que ocorreria por acaso.

Ground Truth	0
LLM	
0.0	115
1.0	182

Fig. 3. Matriz de confusão da amostra de proporção real.

Na amostra obtida por amostragem aleatória representativa da proporção real das classes, observou-se a presença exclusiva de transações legítimas no rótulo de referência, o que inviabiliza uma avaliação efetiva de concordância. Nessa configuração, o coeficiente Kappa resultou em valor nulo (k = 0.0), refletindo a ausência de variabilidade na classe de referência e não um desempenho inadequado do modelo.

Ground Truth	0	1
LLM		
0.0	95	12
1.0	142	47

Fig. 4. Matriz de confusão da amostra de proporção estratificada 80/20.

Já na amostra estratificada, contendo proporções controladas de transações legítimas (80%) e fraudulentas (20%), a matriz de confusão evidenciou a presença de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, permitindo uma avaliação mais informativa. Nessa configuração, o coeficiente Kappa obtido foi $k \approx 0.11$, indicando concordância fraca entre a classificação da LLM e o rótulo real da base.

Esse resultado sugere que, embora a LLM identifique padrões associados a risco financeiro, sua classificação não reproduz diretamente o critério operacional adotado na rotulagem da base IBM. Tal comportamento é esperado, dado que o modelo não foi treinado especificamente para a tarefa de detecção de lavagem de dinheiro nem calibrado com base nos rótulos do conjunto de dados, atuando como um avaliador heurístico baseado em regras implícitas aprendidas durante seu pré-treinamento.

VI. Análise crítica dos resultados

Os resultados obtidos ao longo do experimento fornecem evidências estatísticas consistentes para responder às perguntas de pesquisa propostas, permitindo caracterizar diferenças entre transações legítimas e suspeitas, bem como avaliar relações entre variáveis financeiras e para compreender parte do comportamento da LLM na atribuição de escores de risco.

A análise comparativa entre grupos indicou que transações rotuladas como lavagem de dinheiro apresentam valores significativamente superiores aos das transações legítimas, resultado robusto mesmo diante de distribuições altamente assimétricas e da presença de valores extremos, cenário típico de dados financeiros. A adoção de testes não paramétricos e de intervalos de confiança baseados em reamostragem mostrou-se adequada para garantir validade inferencial nesse contexto.

A análise de correlação revelou que o escore de risco atribuído pela LLM apresenta forte associação com o valor da transação, indicando que o modelo utiliza o montante financeiro como um dos principais critérios na avaliação de risco. Contudo, essa associação não deve ser interpretada como evidência de causalidade real, uma vez que o escore de risco é uma variável derivada e influenciada diretamente pelas instruções fornecidas ao modelo, o que pode introduzir viés no processo de classificação. Por outro lado, não foi observada correlação estatisticamente significativa entre o valor da transação e a frequência de transações da conta, sugerindo independência entre intensidade operacional e volume financeiro no conjunto de dados.

No que se refere à concordância entre a classificação produzida pela LLM e o rótulo real da base de dados, os resultados indicaram concordância fraca quando considerada uma amostra balanceada, evidenciando que o modelo não reproduz diretamente os critérios operacionais adotados na rotulagem da base IBM. Esse comportamento é esperado, dado que a LLM não foi treinada especificamente para a tarefa de detecção de lavagem de dinheiro, atuando como um classificador heurístico baseado em conhecimento geral. A LLM apresenta um viés conservador, classificando uma proporção significativa de transações legítimas como suspeitas, o que sugere alta sensibilidade potencial, porém baixa especificidade.

Por fim, optou-se por privilegiar medidas inferenciais formais, como testes de hipótese, intervalos de confiança e coeficientes de associação, em detrimento de uma abordagem predominantemente exploratória baseada em visualizações gráficas. Essa escolha visou assegurar rigor estatístico e clareza interpretativa, uma vez que os testes empregados capturam de forma adequada as propriedades centrais das distribuições e relações analisadas.

Entre as limitações do estudo, destacam-se o uso de uma LLM genérica, sem ajuste ou treinamento supervisionado

específico para o domínio de detecção de lavagem de dinheiro, bem como o impacto do desbalanceamento das classes na avaliação de concordância. Trabalhos futuros podem explorar estratégias de calibração do modelo, enriquecimento do prompt com variáveis adicionais e comparação com métodos supervisionados tradicionais, visando avaliar ganhos de robustez e desempenho.

REFERENCES

[1] J. C. da Silva Junior, "Lavagem de dinheiro sob a lupa da investigação policial: desafios, táticas e impactos no orçamento público," *Revista Brasileira de Filosofia e História*, vol. 13, no. 1, pp. 2281–2299, Feb. 2024, doi:10.18378/rbfh.v13i1.10357. :contentReference[oaicite:0]index=0

[2] B. Deprez, W. Wei, W. Verbeke, B. Baesens, K. Mets, and T. Verdonck, "Advances in continual graph learning for anti-money laundering systems: a comprehensive review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 17, no. 3, Art. e70040, 2025, doi:10.1002/wics.70040. :contentReference[oaicite:0]index=0

[3] A. Ruchay, E. Feldman, D. Cherbadzhi, and A. Sokolov, "The Imbalanced Classification of Fraudulent Bank Transactions Using Machine Learning," *Mathematics*, vol. 11, no. 13, Art. no. 2862, Jun. 2023, doi:10.3390/math11132862. :contentReference[oaicite:1]index=1

[4] N. Husnaningtyas, G. F. Hanin, T. Dewayanto, and M. F. Malik, "A systematic review of anti-money laundering systems literature: Exploring the efficacy of machine learning and deep learning integration," *JEMA: Jurnal Ilmiah Bidang Akuntansi dan Manajemen*, vol. 20, no. 1, pp. 91–116, Mar. 2023, doi:10.31106/jema.v20i1.20602. :contentReference[oaicite:0]index=0

[5] K. Yang, Z. Zhong, S. Sun, Z. Yu, C. L. P. Chen, and T. Zhang, "LLM40FD: Unlocking the Potential of LLM for Anonymous Zero-Shot Fraud Detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 6, pp. 4606–4619, Dec. 2025, doi:10.1109/TCSS.2025.3563954. :contentReference[oaicite:0]index=0

[6] B. Ramanujam, "Statistical Insights into Anti-Money Laundering: Analyzing Large-Scale Financial Transactions," *Int. J. Eng. Res. Technol. (IJERT)*, vol. 14, no. 4, 2025, doi:10.17577/IJERTV14IS040136.

[7] R. I. T. Jensen, J. Ferwerda, and C. R. Wewer, "Searching for Smurfs: Testing if Money Launderers Know Alert Thresholds," *J. Quant. Criminol.*, 2025, doi:10.1007/s10940-025-09617-7.