Universidade de São Paulo
Escola Politécnica
Programa de Educação Continuada em Engenharia
Especialização em Inteligência Artificial

Matias Cornelsen Herklotz

Identificação de desvios comportamentais no iGaming por meio de aprendizado não supervisionado

São Paulo
2025

MATIAS CORNELSEN HERKLOTZ

Identificação de desvios comportamentais no iGaming por meio de aprendizado não supervisionado

— Versão Original —

Monografia apresentada ao Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para conclusão do curso de Especialização em Inteligência Artificial.

Orientador: Profa. Dra. Larissa Driemeier

São Paulo
2025

Catalogação-na-publicação

Herklotz, Matias Cornelsen
Identificação de desvios comportamentais no iGaming por meio de aprendizado não supervisionado/ M. C. Herklotz – São Paulo, 2025.
53p.

Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1. Apostas online 2. IGaming 3. Adicção comportamental 4. Clusterização 5. Autoencoders.
I. Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia. II.t.

Sumário

Sumário • *ii*

Resumo • *iv*

Abstract • *v*

Lista de Figuras • *vi*

Lista de Tabelas • *vii*

1 Introdução • *1*
- 1.1 Problema • 1
  - 1.1.1 De hábito para adicção • 2
  - 1.1.2 Adicção comportamental • 2
  - 1.1.3 Transtorno de jogo • 3
- 1.2 Objetivo • 4

2 Revisão da literatura • *5*
- 2.1 Adicção comportamental • 5
  - 2.1.1 Dependência química • 6
  - 2.1.2 Ambiente • 7
- 2.2 Métodos supervisionados • 7
- 2.3 Métodos não supervisionados • 8
  - 2.3.1 Agrupamento • 8
  - 2.3.2 Identificação de comportamento anômalo • 9

3 Métodos • *10*
- 3.1 Materiais • 10
  - 3.1.1 Jogo • 10
  - 3.1.2 Conjunto de dados • 11
- 3.2 Processamento • 12
  - 3.2.1 Pré-processamento • 12
  - 3.2.2 Análise exploratória • 12
  - 3.2.3 Agregações e transformações • 12
- 3.3 Modelos • 13
  - 3.3.1 Clusterização • 13
  - 3.3.2 AutoEncoders • 13

4 **Desenvolvimento** • *14*
4.1 Processamento • 14
4.1.1 Pré-processamento • 14
4.1.2 Análise exploratória • 15
4.1.3 Transformações • 16
4.1.4 Agregações • 18
4.1.5 Transformações pós-agregações • 19
4.2 Modelos • 20
4.2.1 Clusterização • 20
4.2.2 AutoEncoders • 27

5 **Resultados e Discussão** • *29*
5.1 Resultados • 29
5.1.1 Análise comparativa entre clusters • 35
5.1.2 Limiar de anormalidade • 37
5.2 Limitações • 40

6 **Conclusão** • *42*

Referências • *43*

Resumo

HERKLOTZ, M. C. *Identificação de desvios comportamentais no iGaming por meio de aprendizado não supervisionado.* 2025. Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia. Universidade de São Paulo, São Paulo, 2025.

Esse trabalho investiga a identificação de desvios comportamentais no contexto do iGaming por meio de técnicas de aprendizado não supervisionado, com o objetivo de analisar o comportamento dos apostadores ao longo do tempo e identificar padrões anômalos. Inicialmente, os apostadores são agrupados de acordo com características comportamentais, o que permite a identificação de um grupo de referência considerado mais estável. A partir desse grupo, são treinados autoencoders para modelar o comportamento esperado, utilizando o erro de reconstrução como medida de distância em relação ao padrão observado. Essa abordagem possibilita a análise diária do comportamento dos indivíduos por meio da quantificação de desvios em relação ao comportamento-alvo. Os resultados indicam que dias com maior erro de reconstrução tendem a apresentar comportamentos mais anômalos, sendo possível empregar um limiar para classificar os dias de aposta. Embora os resultados sugiram que a metodologia proposta seja promissora para a identificação e o acompanhamento de desvios comportamentais, ainda há necessidade de validação adicional, especialmente no que diz respeito à relação entre as anomalias detectadas e os comportamentos problemáticos, bem como à definição e ao uso de limiares ou métodos de distinção de comportamento.

Palavras-chave: Apostas online. IGaming. Adicção comportamental. Clusterização. Autoencoders.

Abstract

HERKLOTZ, M. C. *Identifying behavioral deviations in iGaming through unsupervised learning*. 2025. Monografia (Especialização em Inteligência Artificial) – Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia. University of São Paulo, São Paulo, Brazil. 2025.

This work investigates the identification of behavioral deviations in the context of iGaming through unsupervised learning techniques, with the objective of analyzing gambler's behavior over time and identifying anomalous patterns. Initially, gamblers are grouped according to behavioral characteristics, which enables the identification of a reference group considered more stable. Based on this group, autoencoders are trained to model expected behavior, using reconstruction error as a distance measure with respect to the observed pattern. This approach enables the daily analysis of individual behavior through the quantification of deviations relative to the target behavior. The results indicate that days with higher reconstruction error tend to exhibit more anomalous behavior, making it possible to employ a threshold to classify betting days. Although the results suggest that the proposed methodology is promising for the identification and monitoring of behavioral deviations, further validation is still required, particularly with respect to the relationship between the detected anomalies and problematic behaviors, as well as the definition and use of thresholds or methods for behavioral distinction.

Keywords: Online gambling. IGaming. Behavioral addiction. Clustering. Autoencoders.

Lista de Figuras

4.1 Distribuição das odds e proporção de apostas acima dos limiares. 15
4.2 Distribuição das médias de odds dos usuários e proporção de usuários acima dos limiares. 15
4.3 Distribuição do total de apostas dos usuários e proporção de usuários acima dos limiares. 16
4.4 Distribuição de usuários - melhores 4 dimensões 20
4.5 Distribuição dos clusters de usuários - melhores 3 dimensões. 21
4.6 Distribuição da média dos usuários dentro dos clusters (21 dias). 23
4.7 Distribuição do STD dos usuários dentro dos clusters (21 dias). 25
4.8 Distribuição da inclinação da reta de apostas por cluster e por janela. 26
4.9 Arquitetura do autoencoder. 28

5.1 Histórico de apostas - Usuário 1343144. 31
5.2 Histórico de apostas - Usuário 1357121. 33
5.3 Histórico de apostas - Usuário 1046423. 35
5.4 Distribuição do erro de reconstrução por cluster (escala log). 36
5.5 Distribuição acumulada do erro de reconstrução (ECDF) por cluster. 37
5.6 Histórico de apostas - Usuário 1288121. 38
5.7 Histórico de apostas - Usuário 220550. 39

Lista de Tabelas

4.1 Exemplo de dados após pré-processamento 14
4.2 Exemplo de transformação do valor de uma aposta em uma proporção. . . . 17
4.3 Nº de usuários por Cluster. 21

5.1 Erro de reconstrução do autoencoder por percentil no cluster de referência. . 29
5.2 Top 5 usuários do cluster nºo com maior erro médio. 30
5.3 Top 5 usuários do cluster nºo com menor distancia de erro P80. 34
5.4 Distribuição do erro de reconstrução (MSE) por cluster. 36
5.5 Usuários com aproximadamente 20% de dias atípicos. 38
5.6 Usuários com aproximadamente 50% de dias atípicos. 39

Capítulo 1

Introdução

Cassinos online, sites de apostas esportivas e plataformas de jogos de azar, são todos tipos de sites caracterizados pelo termo *"iGaming"*. [1]

Em comparação com métodos de apostas mais tradicionais, o iGaming se destaca pela acessibilidade, conveniência e flexibilidade (Ghelfi *et al.*, 2024). Um apostador consegue, no conforto de sua casa, apostar o quanto quiser a qualquer hora do dia.

O iGaming tem se tornado cada vez mais popular. Hoje em dia, é comum ver patrocínios de casas de apostas em jogos de futebol, por exemplo. Os brasileiros estão cada vez mais se acostumando com sua presença no dia a dia.

No mês de abril de 2025, brasileiros apostavam até 30 bilhões de reais mensalmente.[2] E a posição do governo brasileiro em relação a tudo isso mostra-se cada vez mais tolerante, com leis mais permissivas e até com o desenvolvimento da *"Bet"* da Caixa Econômica Federal, uma plataforma oficial de apostas esportivas.[3]

1.1 Problema

Por mais que apostar seja uma prática relativamente aceita pela sociedade, diferentes culturas mundiais a veem com diferentes olhos. E isso não é à toa: apostar carrega um potencial destrutivo preocupante para o apostador.

Primeiro, pelo risco financeiro direto: é uma atividade que não tem um custo máximo e depende que o próprio usuário imponha limites a si mesmo.

Segundo, é uma prática que traz ao usuário a excitação de se arriscar e, caso seja vencedor, uma liberação de dopamina que reforça o comportamento (Grant *et al.*, 2010). Um indivíduo começa a apostar de forma recreativa e despretensiosa, e aos poucos vai

[1] JusBrasil. *O que é iGaming e quais são os desafios legais para influenciadores.* 2025.
[2] Reuters. *Brazilians wager up to $5.1 bln a month on online betting, central bank says.* 08 abr. 2025.
[3] TecMundo. *Caixa quer lançar bet própria em novembro e faturar R$ 2,5 bilhões em 2026.* 22 out. 2025.

caminhando para uma relação de **hábito** ou de **adicção comportamental** (Solly *et al.*, 2025).

1.1.1 De hábito para adicção

Para se tornar um hábito, o comportamento deve ser recompensador, ou estar alinhado aos objetivos de longo prazo do indivíduo. Conforme o engajamento aumenta, o resultado do comportamento deixa de ser importante e o processo de tomada de decisão passa a ser cada vez mais automático (Solly *et al.*, 2025; Lamb; Ginsburg, 2018; Güell, 2014).

Um hábito pode ser visto como um passo anterior à adicção comportamental, mas ele não é suficiente por si só (Lamb; Ginsburg, 2018). A transição depende da combinação de diversos fatores, incluindo as características do indivíduo e do comportamento (Güell, 2014).

Um dos fatores contribuintes seria a ausência de restrições ou dificuldades, ou seja, quão fácil é realizar o comportamento. Traçando um paralelo com as substâncias psicoativas, para um fumante, por exemplo, a restrição seria não poder fumar em qualquer lugar ou ter de sair de casa para comprar cigarros. No caso das substâncias ilícitas, as dificuldades incluem a própria ilegalidade e a falta de acesso (Lamb; Ginsburg, 2018).

A presença de restrições não impede necessariamente a formação da adicção, porém, a dificulta. No caso das apostas online, há poucas restrições. Geralmente, o iGaming, como modelo de negócio, busca tornar a jornada de uma aposta a mais simples possível. Seria o equivalente a um fumante ter acesso a infinitos cigarros e poder fumar em qualquer lugar.

Outro fator muito importante é a predisposição do indivíduo. Alguns modelos que buscam explicar a jornada do comportamento adictivo citam algumas características do indivíduo como um elo determinante no engajamento problemático, tais como a compulsividade e a inflexibilidade cognitiva (Solly *et al.*, 2025).

1.1.2 Adicção comportamental

Uma adicção comportamental, ou transtorno de comportamento adictivo, é classificada no ICD-11 sob a categoria *Disorders Due to Addictive Behaviours*,[4] e reflete a perda de controle sobre um determinado comportamento, que persiste independentemente das consequências, podendo ser motivado por recompensas de curto prazo, gatilhos comportamentais ou gatilhos sentimentais, como o desejo de aliviar uma angústia ou impulso (Grant *et al.*, 2010; Solly *et al.*, 2025; Thakur; Kashyap, 2025; Santos Zava; Souza; Messetti, 2024).

[4] Organização Mundial da Saúde. *ICD-11: Disorders Due to Addictive Behaviours.* 2019.

Existe uma gama de atividades com as quais se pode ter uma relação adicta: fazer compras, se alimentar, se exercitar, trabalhar etc. Com a era digital, surgiram novas categorias, como usar internet, jogar videogames, usar redes sociais etc (Thakur; Kashyap, 2025).

Além da persistência do comportamento e da perda de controle, existem outros dois padrões de comportamento que chamam a atenção em indivíduos com esse quadro (Thakur; Kashyap, 2025):

O comportamento é priorizado em relação às atividades básicas do dia a dia, incluindo obrigações e relações interpessoais. Ou seja, consomem o tempo que seria destinado a outras atividades, negligenciando diversos aspectos da vida do indivíduo. Isso pode ser medido, num primeiro momento, pelo número de horas diárias que o indivíduo dedica a esse comportamento.

Outro ponto importante é a tendência do indivíduo de escalar o comportamento, mesmo que muito lentamente. Há uma progressão na intensidade e frequência do comportamento, que também poderia ser identificada ao se analisar o comportamento do indivíduo ao longo do tempo.

1.1.3 Transtorno de jogo

Transtorno de jogo, classificado no ICD-11 como *Gambling Disorder*, no âmbito de *Disorders Due to Addictive Behaviours*,[5] é a adicção comportamental relacionada a apostas, que, além das características já citadas anteriormente, naturais de uma adicção comportamental, também se destaca por afetar diretamente o apostador, tanto financeiramente quanto psicologicamente (Santos Zava; Souza; Messetti, 2024).

O principal problema de um diagnóstico de transtorno de jogo é que essas características nesses apostadores aparecem em diferentes intensidades e em diferentes momentos, e nem todas se apresentam simultaneamente (Solly *et al.*, 2025).

A forma mais fácil de compreender a trajetória do apostador problemático é dividi-la em estágios. No estágio inicial, também chamado de estágio recreativo, o indivíduo pode apostar por diferentes motivos, como influência social de amigos ou colegas, busca de um sentimento de excitação atrelado ao risco, como um certo refúgio emocional, ou então pelo desejo de ganhar dinheiro (Ghelfi *et al.*, 2024).

À medida que o comportamento vai se intensificando, o segundo estágio se manifesta. Também chamado de intermediário, o segundo estágio seria o engajamento perigoso, classificado no ICD-11 como *Hazardous gambling or betting*, [6] nele, há um aumento do risco para o apostador, associado ao contexto das apostas. Por exemplo, a frequência

[5] Organização Mundial da Saúde. *ICD-11: 6C50 – Gambling Disorder.* 2019.
[6] Organização Mundial da Saúde. *ICD-11: QE21 – Hazardous gambling or betting.* 2019.

de apostas ou jogos, o tempo gasto nessas atividades etc. E esse comportamento persiste, mesmo com a ciência do potencial risco. Nesse momento, já seria indicada uma intervenção ou aconselhamento, mas ainda não se pode diagnosticá-lo necessariamente como um distúrbio de jogo (Solly *et al.*, 2025).

O estágio final seria justamente o quadro de transtorno de jogo anteriormente descrito, a etapa final da jornada do apostador problemático, quando ele já pode ser diagnosticado e as soluções mais efetivas já abordam a psicologia e a farmacologia (Solly *et al.*, 2025; Thakur; Kashyap, 2025).

O que dificulta ainda mais é que a duração desses estágios é diferente de pessoa para pessoa: um indivíduo pode passar meses no estágio intermediário, enquanto outro pode alcançar o estágio crítico em uma semana de jogo, por exemplo. E não apenas isso: um indivíduo pode apresentar e intensificar um comportamento problemático em qualquer uma das etapas, sem necessariamente ser diagnosticável (Solly *et al.*, 2025).

1.2 Objetivo

Entende-se que seria ideal impedir ou dificultar que os usuários cheguem ao último estágio do transtorno, a longo prazo, buscando uma harmonia entre o jogador e o jogo.

Como visto anteriormente, diagnosticar o transtorno de jogo nos primeiros estágios é quase inviável. Porém, esses estágios deveriam ser justamente o ponto de ataque, quando o usuário ainda está aprendendo a se relacionar com as apostas e não agravou sua condição.

A proposta deste trabalho, não é identificar os estágios ou entender quando alguém está migrando de um estágio para outro, mas identificar quando alguém está mudando seu comportamento de forma geral, intensificando alguma prática ou fugindo do próprio hábito, independentemente de seu estágio.

Capítulo 2

Revisão da literatura

2.1 Adicção comportamental

Há um grande consenso sobre as definições e implicações das adicções comportamentais.

Geralmente, elas são orientadas por emoções: antes de realizar a ação, o indivíduo sente angústia, ansiedade e excitação (Grant *et al.*, 2010; Solly *et al.*, 2025). Ele sente que **precisa** agir e, ao satisfazer o desejo, experimenta um sentimento de alívio, prazer ou gratificação, que é proporcional ao sentimento anterior, recompensando o comportamento (Grant *et al.*, 2010).

A definição também implica a perda de controle, que seria a falha ao tentar resistir à tentação. Nesse caso, o usuário cede aos sentimentos e se permite fazer o ato (Grant *et al.*, 2010; Solly *et al.*, 2025). Embora o indivíduo se permita repetidas vezes, ele não está necessariamente decidindo isso, pois assim como um hábito é automático, uma adição também é, o que contribui para a perda de controle (Lamb; Ginsburg, 2018).

Por último, as dependências comportamentais são marcadas por suas consequências, que podem variar de acordo com o comportamento, mas geralmente afetam o indivíduo e terceiros nos aspectos sociais, financeiros, emocionais e até físicos (Grant *et al.*, 2010; Solly *et al.*, 2025; Santos Zava; Souza; Messetti, 2024).

Existe um perfil de pessoas que são predispostas a ter problemas com apostas, sendo que, geralmente, se tratam de homens, adolescentes ou jovens adultos (Grant *et al.*, 2010).

A propensão para uma relação problemática pode ter origem em características pessoais e genéticas. Estudos demonstram que pessoas com familiares dependentes de substâncias químicas tendem a gastar mais dinheiro em apostas (Grant *et al.*, 2010; Solly *et al.*, 2025).

2.1.1 Dependência química

A dependência química se assemelha muito a uma adicção comportamental por diferentes motivos.

Emoções como desejo e angústia também estão presentes em dependentes químicos antes do consumo. A dificuldade de resistir, interpretada como perda de controle, também está presente, assim como o sentimento de alívio ou gratificação de uma adicção, também pode ser relacionado ao estado de euforia e excitação provocados pelo uso de substâncias (Grant *et al.*, 2010).

A necessidade de um adicto comportamental de intensificar suas atividades pode ser vista como um comportamento análogo à tolerância de um dependente químico. Se um jogador aposta um valor alto, por exemplo, ele sentirá euforia ou excitação por se arriscar. Porém, quanto mais ele apostar nessas proporções, mais ele se acostumará e, para ter a mesma emoção, precisará aumentar o valor da aposta ainda mais (Grant *et al.*, 2010).

A sensação de que o indivíduo precisa se engajar, motivada por um conjunto de emoções, como a ansiedade e a angústia, é análoga à abstinência química. Ambos parecem cruzar esse limite entre o querer e o precisar (Grant *et al.*, 2010).

Esse fenômeno pode ser explicado pelo fato de que o corpo humano busca estabilidade e, por isso, adapta seus mecanismos fisiológicos às novas práticas do usuário. Os adictos, em geral, entram nesse estado alostático, no qual naturalmente buscam sustentar o que é considerado o novo ”normal”. Porém, a longo prazo, esse estado é, por natureza, insustentável (Koob; Le Moal, 2001).

Essas semelhanças são fundamentais para a elaboração de teorias sobre as dependências comportamentais e sua relação entre si. Ao comparar indivíduos de cada uma delas, observou-se que, em ambas, as pessoas privilegiavam recompensas a curto prazo, tinham dificuldade em analisar ou ponderar risco e recompensa, eram inflexíveis cognitivamente e apresentavam um planejamento geral deficiente (Grant *et al.*, 2010).

No contexto da dependência química, também fala-se em *Unconstrained Demand* (demanda irrestrita), que seria a quantidade de uma substância que seria consumida se houvesse acesso ilimitado a ela.

No que se refere ao transtorno de jogo, o principal limite enfrentado pelo apostador é o valor que ele pode apostar. Portanto, é preocupante que ele alcance o teto ou supere a demanda irrestrita, ou seja, quando não há mais o que apostar. Isso afeta diretamente os indivíduos tanto financeiramente quanto psicologicamente. Com frequência, para sustentar sua adicção, eles se endividam com instituições financeiras, amigos e familiares, e podem até cometer ilegalidades, como fraudes e roubos (Lamb; Ginsburg, 2018; Santos Zava; Souza; Messetti, 2024).

2.1.2 Ambiente

O ambiente desempenha um papel fundamental no desenvolvimento e na manutenção de uma adicção comportamental, abrangendo não apenas o local onde a atividade ocorre, mas também circunstâncias, regras, condições, características e aspectos gerais (Güell, 2014).

O ambiente pode tanto facilitar quanto dificultar a formação de uma dependência ou hábito. Historicamente, quando o uso de cigarros foi restringido em locais públicos, houve uma queda geral no consumo, incluindo o número de cigarros consumidos individualmente por pessoa (Lamb; Ginsburg, 2018).

Ao comparar os apostadores online com os presenciais, por exemplo, nota-se uma grande diferença de intensidade e variabilidade. Os apostadores online apostam por mais sessões e gastam mais do que os apostadores físicos (Ghelfi *et al.*, 2024).

Durante o período de isolamento social decorrente da pandemia de coronavírus, por exemplo, observou-se uma queda nos níveis de jogo problemático em geral, enquanto os aumentos registrados estiveram mais associados a jovens e a altos níveis de sofrimento psicológico (Solly *et al.*, 2025).

O fator social do ambiente também influencia muito o comportamento. Os motivos iniciais que levam um usuário a se engajar em uma prática podem ser oriundos da influência de terceiros ou da busca por aceitação externa (Thakur; Kashyap, 2025).

A aceitação da sociedade em relação à prática, é outro ponto de influência do ambiente social. Relações problemáticas com atividades como a prática de exercícios físicos, por exemplo, são as mais difíceis de serem reconhecidas, pois, para muitos, esses comportamentos são normais (Thakur; Kashyap, 2025). E a popularização da prática a torna mais acessível e fácil de se engajar, impulsionando os possíveis problemas (Santos Zava; Souza; Messetti, 2024).

Por isso, a normalização e a banalização das apostas são pontos delicados que não devem ser negligenciados. E é no ambiente em que as regras são definidas, onde existe o potencial de se conter comportamentos problemáticos.

2.2 Métodos supervisionados

Existem trabalhos cujo foco é classificar ou rotular apostadores com comportamento problemático. Esses trabalhos dependem de exemplos já rotulados, seja de indivíduos que solicitaram a autoexclusão de plataformas de apostas, seja de especialistas que manualmente estudaram e rotularam os dados. Geralmente, atuam nos adictos já estabelecidos, no último estágio.

Apesar disso, estes estudos mostram que, com o histórico de apostas, e informações de depósitos e retiradas, já é possível classificar com sucesso esses apostadores. Por exemplo, Suzuki *et al.* (2019), mostraram com o uso de *shapelets*, que é possível diferenciar de forma consistente jogadores com comportamento problemático daqueles sem indícios de risco com base em seus padrões de comportamento.

Andersson *et al.* (2025) também mostraram que modelos de classificação, treinados com janelas temporais curtas (30, 60 e 90 dias) mantêm desempenho preditivo comparável ao uso de históricos completos, sugerindo que sinais relevantes de risco emergem em horizontes temporais reduzidos.

2.3 Métodos não supervisionados

2.3.1 Agrupamento

Existem trabalhos onde o foco é agrupar e estudar os grupos de apostadores. Esses trabalhos por sua vez, não dependem de um rótulo para seu desenvolvimento, mas trabalhos como o de Challet-Bouju *et al.* (2020), demonstram não apenas que é possível agrupar com sucesso os apostadores, mas também que o cluster considerado como o de problemáticos, era o único que continha pessoas que solicitaram a autoexclusão de plataformas.

Outra abordagem para agrupar os usuários foi a de Peres *et al.* (2021), que usaram séries temporais para agrupar o comportamento de indivíduos, similar à abordagem dos *shapelets* de Suzuki *et al.* (2019). Esses métodos se mostraram adequados, pois, conforme visto anteriormente, as transições de comportamento ocorrem ao longo do tempo, em momentos e ritmos diferentes. Dessa forma, analisam e comparam comportamentos de uma maneira dinâmica.

Alguns trabalhos, como os de Challet-Bouju *et al.* (2020) e Auer e Griffiths (2022), também consideraram, além do comportamento do indivíduo (ou seja, valores apostados, frequência e intensidade), suas características demográficas (idade, gênero, entre outras). Com sucesso, esses modelos mostraram que essas informações também são relevantes para agrupar os jogadores e seus comportamentos.

Challet-Bouju *et al.* (2020) também mostraram que o grupo considerado problemático representava uma parcela pequena de jogadores, aproximadamente 5%. Dessa forma, pode-se concluir que jogadores com comportamento problemático geralmente constituem uma minoria e podem ser tratados como uma anomalia.

Do ponto de vista teórico, agrupar os apostadores em n grupos é mais adequado do que classificá-los como problemáticos ou não problemáticos. Conforme visto anteriormente, existem diferentes estágios de risco e, além disso, dentro de cada estágio, existem grupos que se comportam de maneiras distintas.

2.3.2 Identificação de comportamento anômalo

Existem também trabalhos focados na detecção de anomalias, como o de Lajčinová, Gall e Michal (2023), que usam *autoencoders* para identificar elementos anômalos. São criadas métricas relacionadas a comportamentos dentro das apostas, bem como agregações dentro de janelas temporais.

O modelo é treinado com o objetivo de reconstruir esse recorte temporal do comportamento de uma pessoa. A partir disso, estudaram-se os 5% com maior erro de reconstrução e concluiu-se que esse mesmo grupo tinha um comportamento mais problemático, com maior número de pedidos de aumento de limite, de logins etc (Lajčinová; Gall; Michal, 2023).

A ideia de que o comportamento anômalo é o problemático, é corroborada pelas conclusões de Challet-Bouju *et al.* (2020), anteriormente discutidas.

Capítulo 3

Métodos

3.1 Materiais

3.1.1 Jogo

Para estudar o comportamento dos usuários no iGaming, selecionou-se o jogo chamado *"Crash"*, um tipo de jogo de azar online que ocorre em rodadas simultâneas, com a participação de vários jogadores. Antes do início de cada rodada, os participantes definem o valor de suas apostas. Depois, há um multiplicador de ganho que começa em 1,0x e sobe gradualmente ao longo da rodada.

O objetivo do jogador é decidir o momento de encerrar sua aposta, saindo da rodada antes que o jogo pare de forma imprevisível. Se o jogador encerrar a aposta a tempo, receberá o valor apostado multiplicado pelo fator exibido naquele instante, mas se o jogo parar antes da retirada, a aposta será perdida.

Não há um valor máximo para o multiplicador, e nem um mínimo para a parada, embora, na prática, os multiplicadores mais comuns estejam entre 1,0x e 2,0x, pois, para o cassino obter lucro, a chance do usuário duplicar seu dinheiro precisa ser obrigatoriamente menor que 50%.

O Crash é particularmente relevante para este trabalho, pois jogadores com perfil de transtorno de jogo tendem a apresentar desvantagem em tarefas que envolvem a análise de risco e recompensa (Grant *et al.*, 2010). Como o multiplicador cresce de forma exponencial, o jogo estimula a tomada de risco contínua.

Por exemplo, se um jogador encerra a rodada em 1,5x, mas observa que o multiplicador chegou a 200x, ele pode sentir que "perdeu a chance de ganhar", mesmo tendo obtido lucro.

Como é o usuário quem determina as circunstâncias da aposta, muitos adotam estratégias específicas, que transmitem uma sensação de controle e segurança. Mesmo

que os resultados continuem aleatórios, existe um grande potencial para a diversidade comportamental.

Além disso, o jogo permite observar padrões coletivos de comportamento, pois as apostas ocorrem em grupo e os resultados são compartilhados em tempo real. Dessa maneira, é possível identificar mudanças no comportamento coletivo, como uma maior propensão ao risco após rodadas de perdas ou ganhos consecutivas, ampliando o potencial analítico para o estudo do comportamento dos jogadores em ambientes de risco social.

3.1.2 Conjunto de dados

O *dataset* ou conjunto de dados, está publicamente disponível no Kaggle,[1] e trata-se de uma coleta de dados históricos de apostas do jogo crash dentro do site *bc.game*, onde os jogadores realizam suas apostas com criptomoedas.

O período de coleta de dados foi de 8 a 30 de setembro de 2020, com um total de 21 dias e 7 horas de apostas. Apesar do curto intervalo, o conjunto de dados conta com informações de 50 mil usuários, 71 mil partidas e um total de quase 26 milhões de apostas.

O dataset está dividido em duas tabelas: uma com informações das partidas e outra com informações das apostas individuais.

Na tabela de partidas, as informações mais importantes são:

- O código de identificação daquela partida; *game_id*.
- O valor do multiplicador no qual a partida terminou; *max_rate*.
- A data e a hora da partida; *timestamp*.

Na tabela de apostas, as informações mais importantes são:

- O código de identificação da partida; *game_id*.
- O código de identificação do usuário; *user_id*.
- O código de identificação da aposta; *bet_id*.
- A modalidade de jogo; *game_type*.
- O valor do multiplicador no momento em que o usuário saiu da partida; *odds*.
- Se a criptomoeda utilizada na aposta pode ser convertida em dólares; *fiat_is_valuable*.
- Em dólares, qual o valor apostado; *fiat_bet_amount*.

[1] Kaggle. *bc.game Crash Dataset [Historical]*. 2022.

- Em dólares, qual o valor do lucro; *fiat_profit_amount*.

A *odd*, ou multiplicador da aposta, não representa diretamente a chance de vitória na partida; seu valor representa exclusivamente o quanto o valor apostado será multiplicado caso a vitória ocorra, embora o valor da odd seja na pratica, inversamente proporcional à chance de vitória, podendo então, ser interpretado como um fator de risco.

3.2 Processamento

3.2.1 Pré-processamento

Nessa etapa, buscou-se garantir a integridade dos dados, filtrando-se colunas menos pertinentes e limpando-se possíveis erros ou inconsistências.

O objetivo foi organizá-los, padronizá-los e estruturá-los, de modo a garantir coerência e permitir uma análise subsequente. Os pontos de atenção incluíram a forma como os dados foram carregados, seus tipos e características e como foram usados.

3.2.2 Análise exploratória

A análise visou compreender o comportamento dos dados de maneira geral e identificar padrões ou tendências. Dado um conjunto de indivíduos, buscou-se entender:

a) O que seria considerado normal ou anormal e quais são as características que compõem um determinado indivíduo ou grupo.

b) Quais diferentes granularidades poderiam ser construídas e quais diferentes insights elas podem proporcionar.

c) Quais métricas importantes estão ausentes, e devem ser construídas.

3.2.3 Agregações e transformações

Dado que se tem uma granularidade, colunas de identificação e informações de tempo, foram analisados quais tipos de agregações fariam sentido para o trabalho e quais poderiam ajudar a compreender o comportamento de um indivíduo ao longo do tempo.

Também buscou-se transformar e combinar colunas de modo a ponderar corretamente a importância de diferentes medidas e informações, bem como construir novas variáveis com potencial explicativo.

Por fim, foi necessário ajustar e preparar os dados para as tarefas seguintes, buscando-se chegar ao final, a um conjunto de dados pronto para treinamento.

3.3 Modelos

Para mensurar quando alguém apresenta um comportamento potencialmente problemático, adotou-se uma estratégia em duas etapas.

Primeiro, os usuários foram agrupados para se identificar qual conjunto representava o comportamento considerado estável ou "desejado".

Em seguida, treinou-se um segundo modelo capaz de aprender e reconhecer o padrão desse grupo. A ssim, dado um dia de apostas, o modelo deveria indicar o quão próximo esse dia estaria do comportamento de referência.

3.3.1 Clusterização

Para formar os grupos, foram utilizados métodos tradicionais de aprendizado não supervisionado, não para se prever risco diretamente, mas para separar perfis de comportamento. Este método permite identificar um cluster cuja dinâmica de apostas seja mais estável ao longo do tempo, servindo como âncora para a etapa seguinte.

3.3.2 AutoEncoders

Após identificado o cluster de referência, treinou-se um autoencoder usando-se somente informações dos usuários pertencentes a esse grupo. Autoencoders aprendem a reconstruir padrões característicos dos dados de entrada; portanto, esperava-se que reconstruíssem bem os dias de aposta do grupo estável e apresentassem um erro maior quando recebessem dados de perfis diferentes.

Na prática, o autoencoder mostrou-se sensível à inicialização dos pesos e à ordem de apresentação dos dados durante o treino.

Para mitigar essa variabilidade e aumentar a robustez da detecção de anomalias, adotou-se uma abordagem de *ensemble*, treinando-se múltiplos autoencoders independentes e combinando seus erros de reconstrução.

O erro de reconstrução passou a funcionar como uma medida contínua de anomalia: quanto maior o erro, maior a distância entre o comportamento observado e o comportamento considerado estável.

E a dispersão interquartil do erro de reconstrução entre os autoencoders, passou a funcionar como um indicativo de quão concordantes eles são entre si: Quanto menor a distância entre os quartis, maior a certeza do erro calculado.

Capítulo 4

Desenvolvimento

4.1 Processamento

4.1.1 Pré-processamento

Iniciou-se o projeto com a importação dos dados, que foram unidos em uma única tabela pelo *game_id*.

As apostas que não tinham um valor equivalente em dólar foram removidas, pois buscou-se comparar diferentes apostas de uma mesma pessoa ou grupo e para isso, era imprescindível que elas estivessem na mesma moeda.

Foram perdidos cerca de 18 mil usuários no processo, que representavam apenas 9% do total de apostas, reduzindo-se o número total para aproximadamente 23 milhões.

Nas apostas em que o jogador perdeu o jogo, a coluna odd assumia um valor zero, nesses casos, substituiu-se o valor pelo valor do multiplicador máximo da rodada, ou seja, até que ponto a pessoa esperou para perder.

Além disso, existiam alguns jogos em que a modalidade de jogo não era a normal. Na prática, a única diferença era que o usuário decidia, antes da rodada, em que momento gostaria de sair. Então, se a modalidade é *red*, a odd dele passa a ser 1,96; se é *green*, 2; e se é *yellow*, 10. Então também substituiu-se o valor nesses casos.

game_id	user_id	bet_id	odds	bet_amount	profit_amount	datetime
2828375	1408934	210854432	2.00	0.1	0.1	2020-09-08...
2828375	367240	210854355	6.00	1.25e-11	6.27e-11	2020-09-08...
2831698	44	212042687	1.03	3.4e-5	1e-6	2020-09-09...
2828664	5588	12159782	10.00	7.09	-7.09	2020-09-09...

Tabela 4.1: Exemplo de dados após pré-processamento

4.1.2 Análise exploratória

A primeira característica analisada foi a das odds. Conforme mencionado anteriormente, a odd pode ser interpretada como um fator de risco, ou seja, a chance do usuário ganhar aquela aposta. Pode ser visto também como o tempo em que o usuário esperou na rodada.

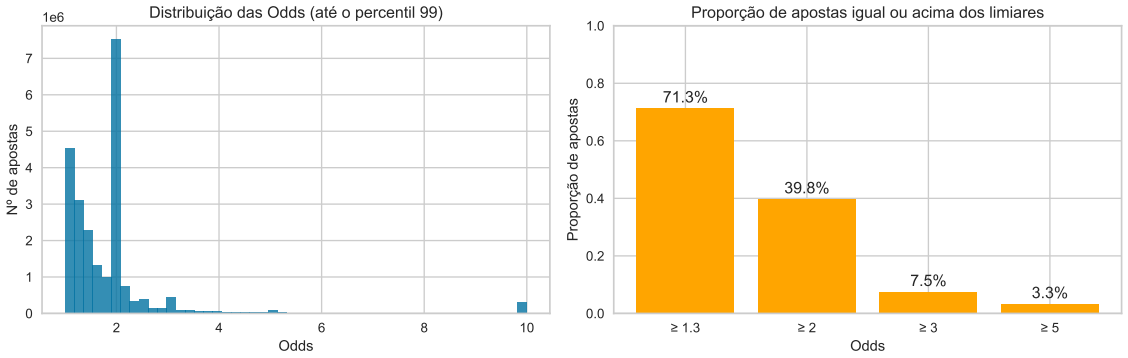

Figura 4.1: Distribuição das odds e proporção de apostas acima dos limiares.

A partir dos gráficos da figura 4.1 é possível fazer algumas observações relevantes. Primeiro, que 30% das apostas possuem odds abaixo de 1,3, o que pode-se considerar como apostas conservadoras, com baixo risco. Segundo, 60% das apostas possuem odds abaixo de 2, são apostas mais comuns, ainda relativamente conservadoras, com chance de ganho maior que 50%. Por fim, apenas 7,5% das apostas possuem odds iguais ou superiores a uma odd 3,0, sendo estas apostas aquelas em que a chance de ganhar é menor que 33%.

Ou seja, a esmagadora maioria das apostas pode ser considerada normal. Existem apostas com odds de 20, 50, 100, entre outras, mas elas representam uma minoria.

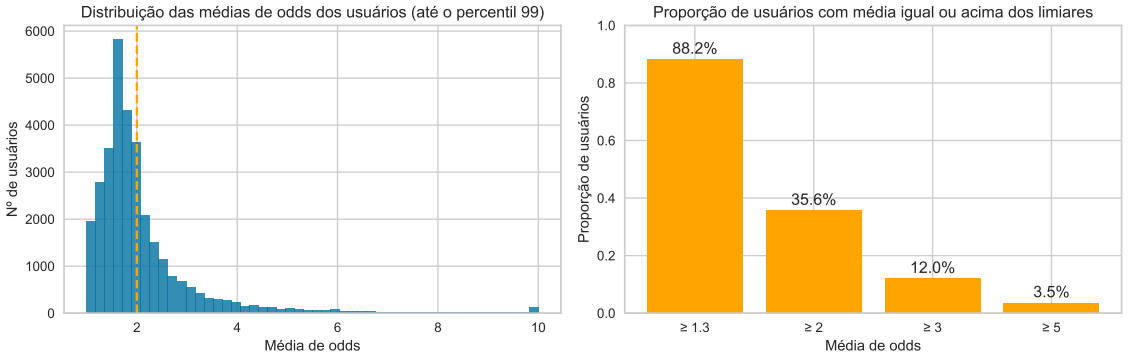

Figura 4.2: Distribuição das médias de odds dos usuários e proporção de usuários acima dos limiares.

Ao se comparar as figuras 4.1 e 4.2, nota-se uma semelhança na distribuição da média de odds das apostas e da média de odds dos usuários, porém, há uma proporção menor de usuários com média de odds abaixo de 1,3 do que de apostas com odds na mesma faixa,

e uma proporção maior de usuários com média de odds iguais ou maiores que 3,0 do que de apostas com odds na mesma faixa.

Isso pode indicar que as pessoas não apostam sempre na mesma odd, e que esse comportamento é dinâmico, ou então que grupos distintos, de conservadores e arrojados, fazem apostas em quantidades diferentes.

Outra análise que pode-se fazer é sobre o número de apostas totais de cada usuário.

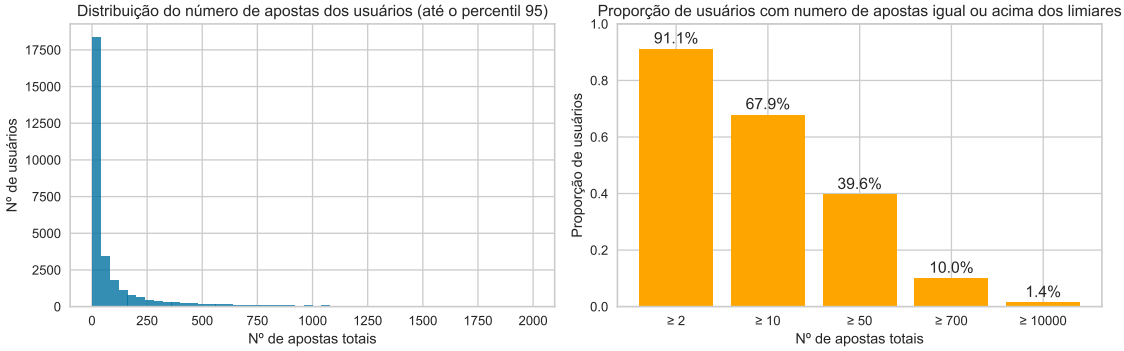

Figura 4.3: Distribuição do total de apostas dos usuários e proporção de usuários acima dos limiares.

Ao analisar a figura 4.3, nota-se que existe um número considerável de jogadores que apostaram apenas uma vez ao longo da coleta. Aproximadamente um terço não chegou a apostar mais de dez vezes, e mais da metade está abaixo de 50 apostas totais.

Nessa etapa, pode-se argumentar que existe um número mínimo necessário de apostas para considerarmos um usuário, mas entende-se que um comportamento de baixíssima intensidade ainda é um comportamento que deve ser analisado. Portanto, optou-se por não remover nenhum grupo.

Ao se dividir o número total de apostas pelo total de usuários, tem-se algo em torno de 700 apostas por jogador. No entanto, o que se vê no gráfico, é que a parcela de jogadores com mais de 700 jogos representa apenas 10% do total de usuários, o que indica que existe um grupo pequeno de jogadores que representa um grande conjunto de apostas.

4.1.3 Transformações

Valor apostado

Para se considerar uma aposta arriscada, deve-se levar em conta dois fatores principais: a chance de vitória da aposta, citada anteriormente como a odd da aposta, e o valor apostado. Os dois se complementam e, quando um está ausente, o risco da aposta é suavizado.

No entanto, há um motivo para não se ter analisado o valor apostado antes: um valor alto ou baixo de uma aposta é relativo. O que é muito para uma pessoa, pode ser pouco para outra.

Para cada usuário, criou-se uma coluna que representa a proporção de cada aposta em relação ao valor que ele costuma apostar. Para isso, calculou-se uma média móvel exponencial das apostas anteriores, em que as apostas mais recentes têm peso maior e as mais antigas vão perdendo importância. Em seguida, dividiu-se o valor da aposta atual por essa média, obtendo-se assim uma medida de quão acima ou abaixo do padrão aquela aposta está.

bet_id	user_id	bet_amount	bet_proportion
12159565	5588	3.54	1.00
12159555	5588	3.54	1.00
12159563	5588	1.77	0.50
12159602	5588	14.18	4.02
12159647	5588	28.36	7.80
12159688	5588	14.18	3.65

Tabela 4.2: Exemplo de transformação do valor de uma aposta em uma proporção.

Essa abordagem permite tanto comparar o comportamento de indivíduos com diferentes realidades financeiras quanto analisar a evolução do comportamento de um mesmo usuário ao longo do tempo. Em um cenário ideal, a proporção de apostas de alguém deveria se manter o mais próximo possível de 1.

Com a suavização, quando a pessoa mudar seu padrão de apostas, a proporção se adaptará ao novo normal, se aproximando lentamente do 1. Porém, se a mudança for muito abrupta ou houver mudanças constantes, a proporção não conseguirá se adaptar, o que destacará um comportamento anormal. Ao se analisar a média da proporção da tabela *4.2*, por exemplo, chega-se a um valor de 2,99, indicando que o indivíduo em questão intensificou seu comportamento.

Peso do risco

Outra transformação que foi aplicada foi o uso do log tanto da odd quanto da proporção do valor apostado. O aumento de risco ao passar de uma odd de 2x para 10x é significativo. Já incrementos como de 10x para 100x ou de 100x para 1000x representam variações relativas menores.

O mesmo raciocínio vale para o valor apostado: apostar 10 vezes o valor habitual é um sinal relevante, enquanto apostas 1000 vezes maiores podem simplesmente refletir que o valor habitual era muito baixo.

O log, comprime a escala e atribui pesos mais proporcionais às variações relativas, evitando que valores extremos dominem a análise.

4.1.4 Agregações

Agregação temporal

Embora seja interessante trabalhar com a granularidade das apostas, isso é custoso. Um indivíduo pode fazer milhares de apostas em um único dia, e como busca-se padrões em uma janela maior, é mais fácil agregar o comportamento de uma pessoa em um período de tempo.

Então, ao invés de analisar todas as apostas de uma pessoa, analisou-se o resumo dos dias em que ela apostou. Criou-se, então, uma tabela com cada dia de aposta de cada jogador e foram feitas as seguintes agregações:

- Número total de apostas feitas; *n_bets*.
- Número total de sessões de jogo, entende-se que duas apostas pertencem à mesma sessão quando o intervalo entre elas é menor que 60 minutos.; *n_sessions*.
- Número total de horas ativas; *n_hours*.
- Média de odds; *odds_log_mean*.
- Desvio padrão de odds; *odds_log_std*.
- Média da proporção de apostas; *bet_prop_log_mean*.
- Desvio padrão da proporção de apostas; *bet_prop_log_std*.
- O quanto a hora com a maior aposta representa do total apostado no dia, em porcentagem; *bet_top1hour%*.
- O quanto as duas horas com a maior aposta representam do total apostado no dia, em porcentagem; *bet_top2hour%*.
- O quanto as quatro horas com a maior aposta representam do total apostado no dia, em porcentagem; *bet_top4hour%*.
- A inclinação da reta de apostas, mostra qual foi a tendência dos valores apostados: se diminuíram, aumentaram ou se mantiveram constantes ao longo do dia; *slope_bet_day*.
- A inclinação da reta de odds, mostra qual foi a tendência das odds: se diminuíram, aumentaram ou se mantiveram constantes ao longo do dia; *slope_odd_day*.

Para o total de apostas no dia, também foi feito o mesmo tratamento dado ao valor apostado. Novamente, calculou-se uma média móvel exponencial do número de apostas anteriores.

Dessa forma, não foram separadas as pessoas que apostam mais ou apostam menos, mas sim aquelas que mais ou menos alteram o número de apostas realizadas.

Agregação de usuários

No entanto, apenas analisar os dias de forma individual não foi suficiente, pois desejava-se acompanhar a mudança de uma pessoa ao longo do tempo. Para isso, fez-se uma segunda agregação, dessa vez sobre todos os dias de um mesmo usuário.

Foram selecionados quatro intervalos de tempo: 2, 7, 14 e 21 dias, e para cada uma das características de um dia, calculou-se a média e o desvio padrão dentro da janela de tempo. Para a feature *n_sessions*, por exemplo, foram criadas as seguintes:

- *n_sessions_mean_2d*.
- *n_sessions_std_2d*.
- *n_sessions_mean_7d*.
- *n_sessions_std_7d*.
- *n_sessions_mean_14d*.
- *n_sessions_std_14d*.
- *n_sessions_mean_21d*.
- *n_sessions_std_21d*.

Chegou-se então a uma tabela contendo uma linha por usuário, com as médias e os desvios-padrões de todas as características de um dia.

4.1.5 Transformações pós-agregações

O conjunto de dados de usuários ficou muito largo, mesmo após se remover as colunas de identificação, totalizando 97 colunas. Para mitigar esse problema, aplicou-se o PCA, uma técnica para reduzir a dimensionalidade dos dados. Como o PCA é sensível a valores extremos e o conjunto de dados os contém, aplicou-se o *RobustScaler*, uma técnica de padronização, antes da decomposição, a fim de equilibrar melhor as features dos dados.

Após escalar os dados, buscou-se com o PCA, 98% de explicabilidade, reduzindo-se o número de colunas de 97 para apenas 12 e perdendo-se apenas 2% de informação. Como

muitas colunas foram construídas, era de se esperar que houvesse uma grande correlação entre a maioria.

Mesmo após todo o tratamento anterior, ainda existiam usuários extremamente atípicos. Como pretende-se agrupar os usuários, para evitar que casos raríssimos distorcessem a formação dos grupos, aplicou-se um clipping dos valores projetados nos componentes principais, limitando-os aos percentis 0,1 e 99,9, com um mínimo impacto na distribuição geral dos dados.

A Figura 4.4 apresenta a distribuição dos usuários após a aplicação do PCA, considerando as quatro primeiras dimensões.

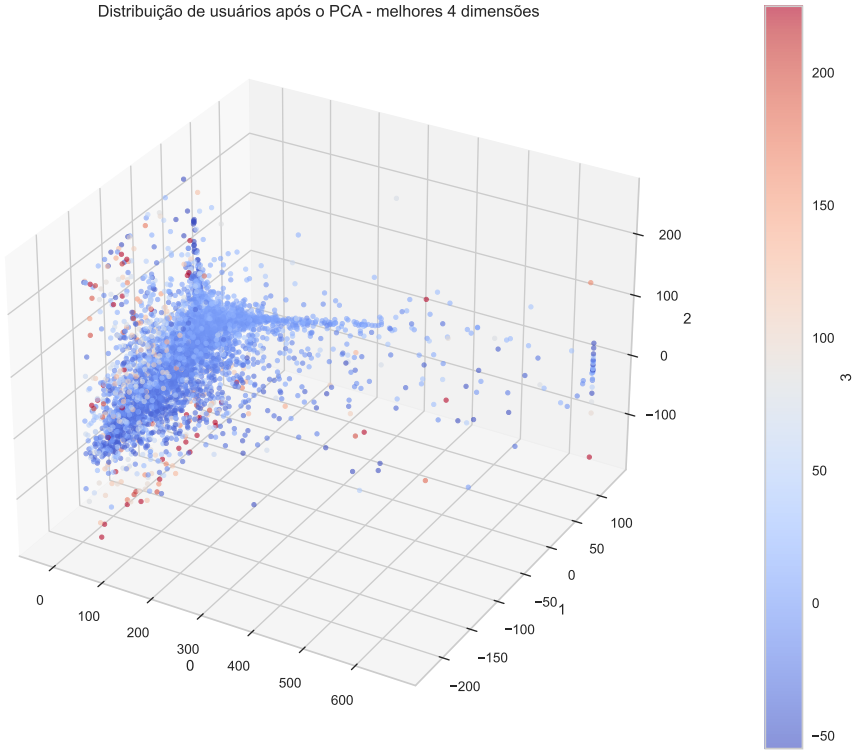

Figura 4.4: Distribuição de usuários - melhores 4 dimensões

4.2 Modelos

4.2.1 Clusterização

A estratégia para definição de modelos e hiperparâmetros de clusterização foi mais experimental e iterativa: após cada clusterização, estudou-se os clusters e suas características para entender se explicavam o que consta na literatura.

No final, o que mais atendeu foi um K-Means com seis clusters, conforme disposto na tabela 4.3

Cluster	Total users	Total gamble days
Cluster nº0	25746	71316
Cluster nº1	1957	20379
Cluster nº2	413	1951
Cluster nº3	3442	30635
Cluster nº4	456	5989
Cluster nº5	88	374

Tabela 4.3: Nº de usuários por Cluster.

Por mais que os clusters tenham tamanhos bem distintos, isso não necessariamente foi um problema. Como visto anteriormente, uma boa parcela dos usuários tinha um comportamento conservador, enquanto uma minoria tinha um comportamento mais intenso, considerando apenas as odds e o número de apostas.

A Figura 4.5 apresenta a distribuição dos usuários após a separação pro grupos, considerando as três primeiras dimensões.

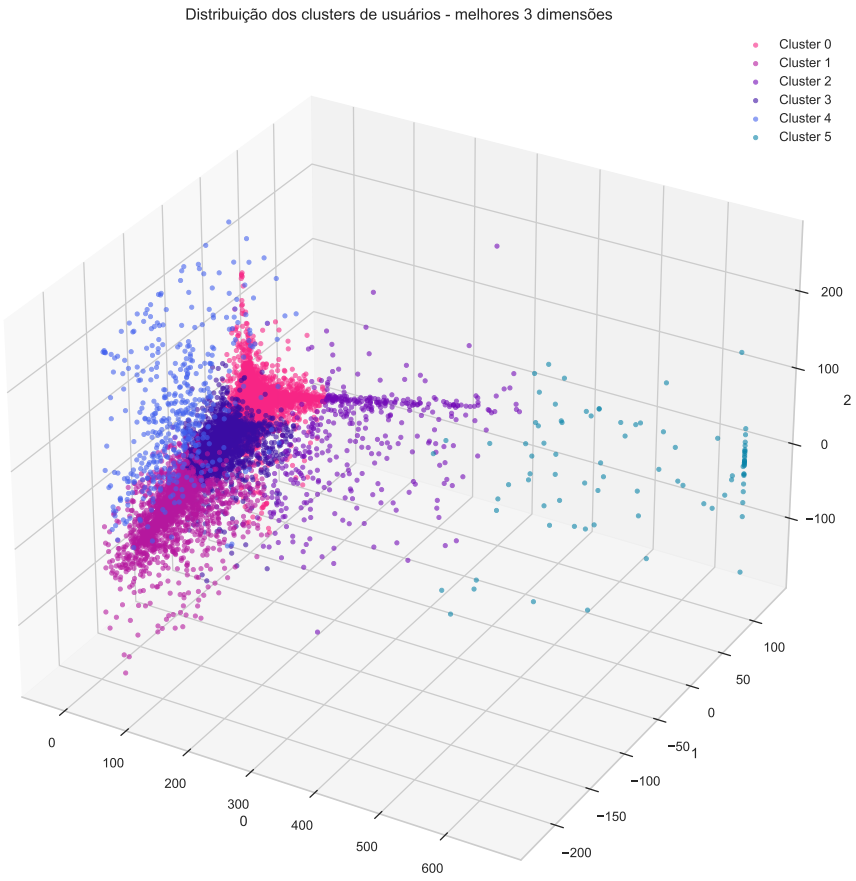

Figura 4.5: Distribuição dos clusters de usuários - melhores 3 dimensões.

Para entender se os clusters estavam bem divididos, analisou-se a diferença de média das características entre os clusters, conforme disposto na figura 4.6.

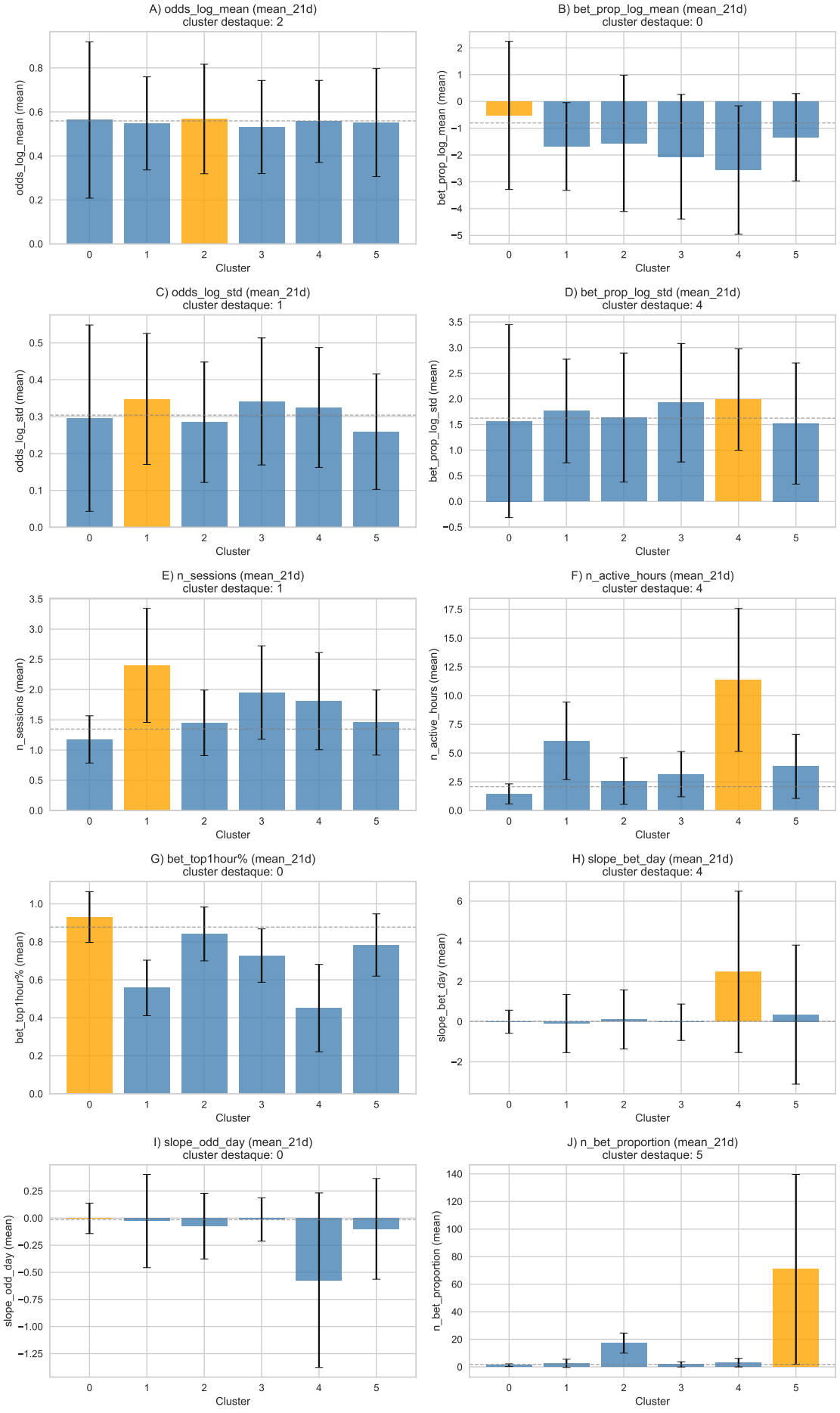

Figura 4.6: Distribuição da média dos usuários dentro dos clusters (21 dias).

Ao se observar os gráficos na figura 4.6, percebe-se que o **Cluster nº0** se aproxima mais de um comportamento constante. Por exclusão, todos os outros apresentam ao menos um tipo de comportamento anômalo, o que significa que:

- **Cluster nº1**: Se destaca pelo número elevado de sessões (E);
- **Cluster nº2**: Se destaca pelo número elevado de proporções de aposta (J);
- **Cluster nº3**: Ele não necessariamente tem um comportamento estranho, mas no geral é mais ativo que o cluster nº0;
- **Cluster nº4**: Se destaca principalmente pelas inclinações de reta tanto do valor apostado quanto da odd, muda bastante seu comportamento ao longo do dia (H, I). Também tem um número altíssimo de horas ativas (F);
- **Cluster nº5**: Chama a atenção pelo número altíssimo de proporção do número de apostas diárias. São pessoas que, ao longo do tempo, aumentaram muito sua atividade (J);

Analisou-se também a diferença de média dos desvios-padrão das características entre os clusters, ou seja, o quanto os próprios usuários mudam de comportamento ao longo do tempo avaliado conforme a figura 4.7.

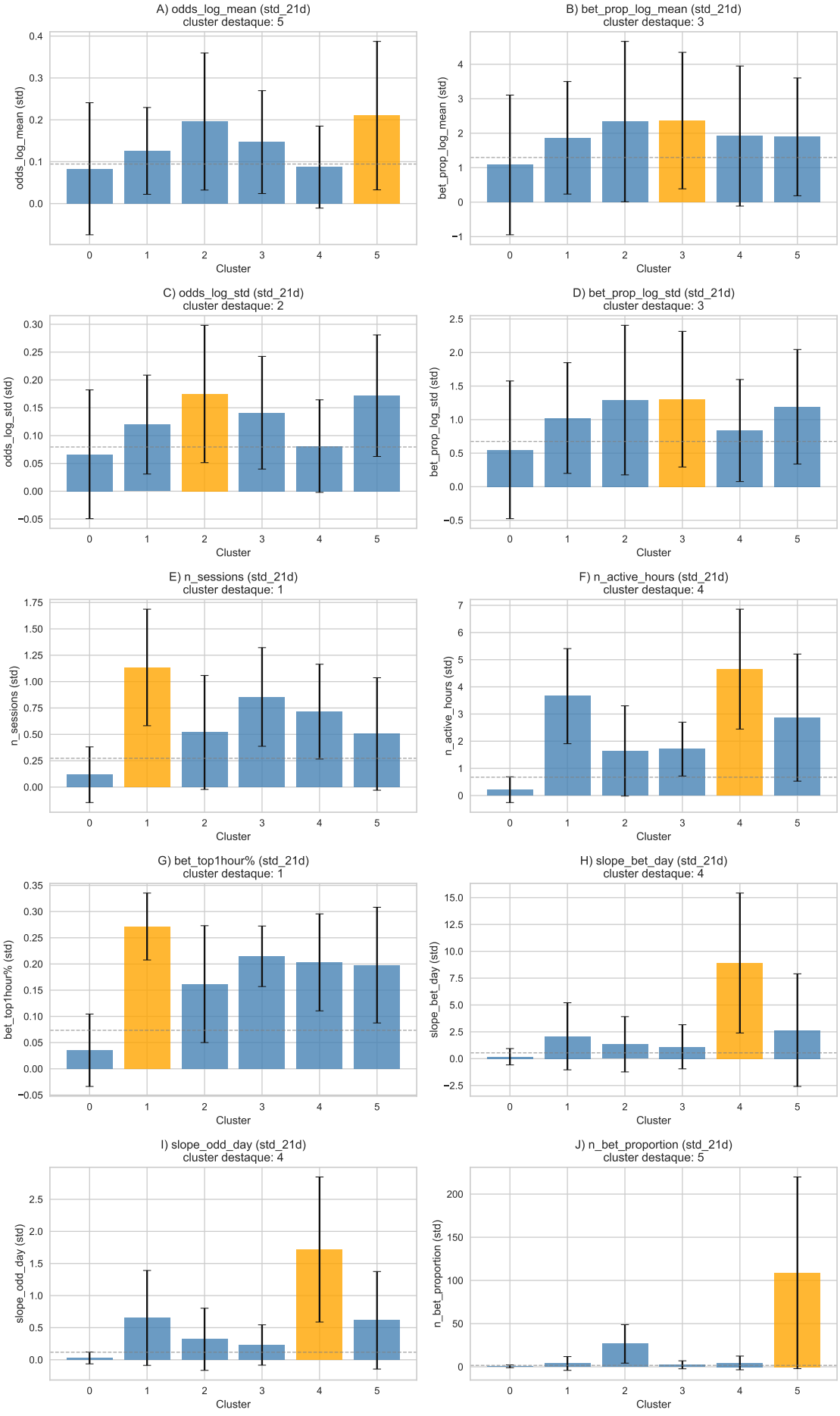

Figura 4.7: Distribuição do STD dos usuários dentro dos clusters (21 dias).

Pode-se ver na figura 4.7, principalmente que os desvios padrões acabam acompanhando as características citadas na análise das médias, mas o ponto mais forte é que o **cluster nº0** é o que se mantém mais constante, com comportamento mais regular.

A análise dos gráficos da figura 4.8, que apresentam as diferenças entre as médias das janelas temporais, indica que, de modo geral, entre os clusters, não há variações expressivas entre as janelas. Ainda assim, a diferença mais evidente ocorre entre os clusters nº4 e nº5, particularmente no que se refere à inclinação da reta de apostas.

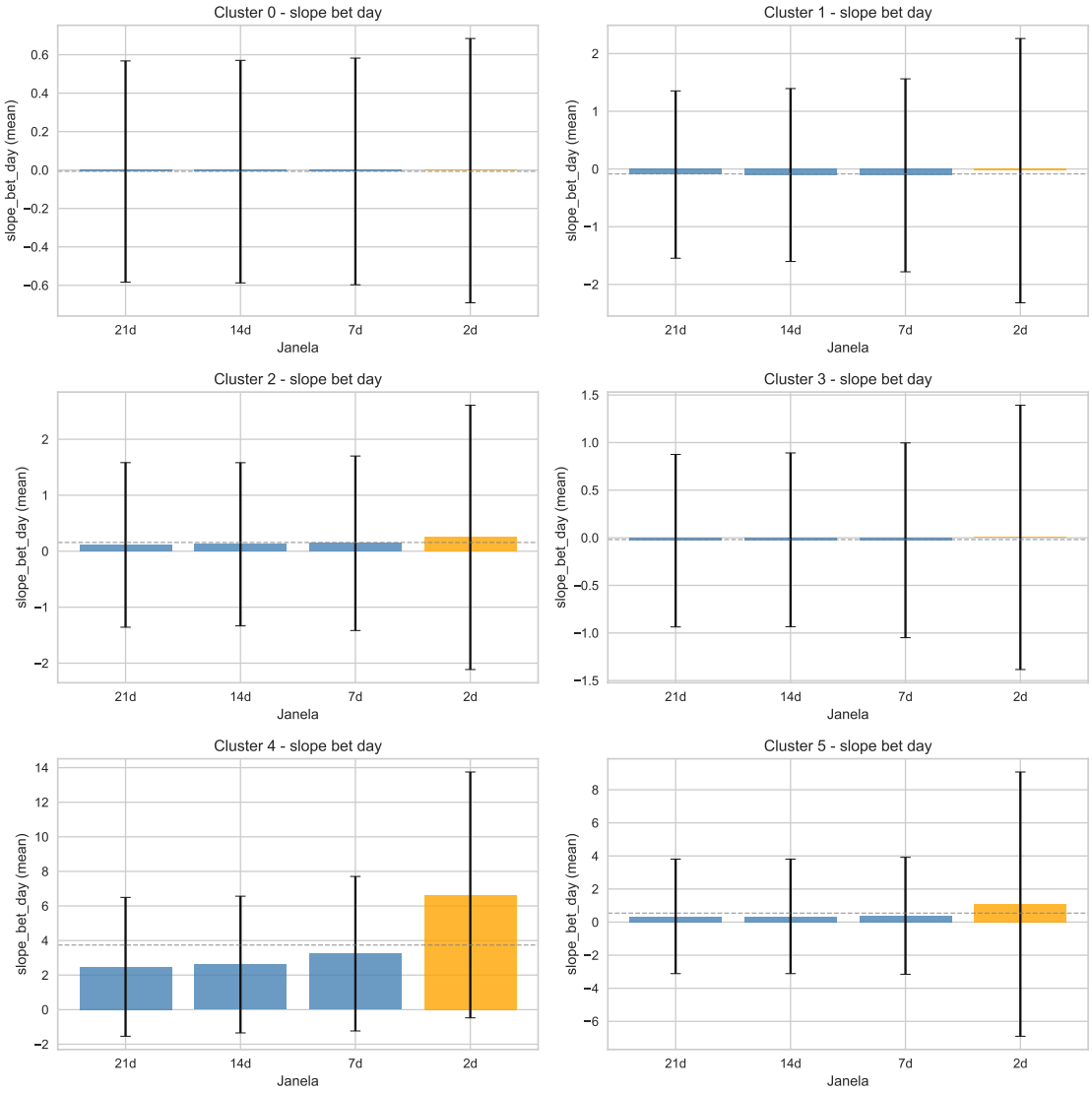

Figura 4.8: Distribuição da inclinação da reta de apostas por cluster e por janela.

Pode-se pensar que o modelo falhou em diferenciar as janelas, mas isso não é necessariamente algo ruim. Primeiro, as janelas podem não representar tempo suficiente para uma mudança de comportamento; elas refletem simplesmente o tempo que se dis-

punha de informação. Segundo, muitas das features já carregam mudanças temporais, tornando a presença de janelas possivelmente redundante.

Pode-se, portanto, interpretar que o Cluster nº0 representa o comportamento que se deseja incentivar em um jogador, não se tratando necessariamente de um perfil com baixo volume de apostas, mas de um comportamento estável, característica central que se busca identificar nesta análise.

No final, buscou-se entender se fazia sentido fazer uma subclusterização do cluster nº0, mas isso se mostrou inviável, pois esses subclusters apresentaram praticamente nenhuma distinção.

4.2.2 AutoEncoders

Para capturar o comportamento dos usuários do cluster nº0, foram construidos sete autoencoders independentes, utilizando todos os dias de apostas desses jogadores.

Foi feita uma nova rodada de pré-processamento para preparar os dados dos dias para o treinamento, as colunas de identificação foram removidas, e novamente normalizou-se os dados.

Como os dados possuem 12 colunas, a arquitetura de cada autoencoder foi composta da seguinte estrutura: Camada de entrada com 12 neurônios; Uma camada intermediária com 8 neurônios; O espaço latente com 4, mais uma intermediária com 8 e a camada de saída com 12, conforme ilustra a figura 4.9.

Os autoencoders foram treinados por 50 épocas em GPU. Usou-se MSE como perda de reconstrução e Adam com taxa de aprendizado 10^{-3}.

A mediana dos sete erros de reconstrução foi utilizada como forma de agregação dos diferentes autoencoders, garantindo que o erro é decidido pela maioria e que os autoencoders com uma decisão muito diferente são desprezados.

Ainda se pode recuperar a indecisão dos autoencoders com a distribuição interquartil, que mostra justamente a diferença entre os erros dos modelos.

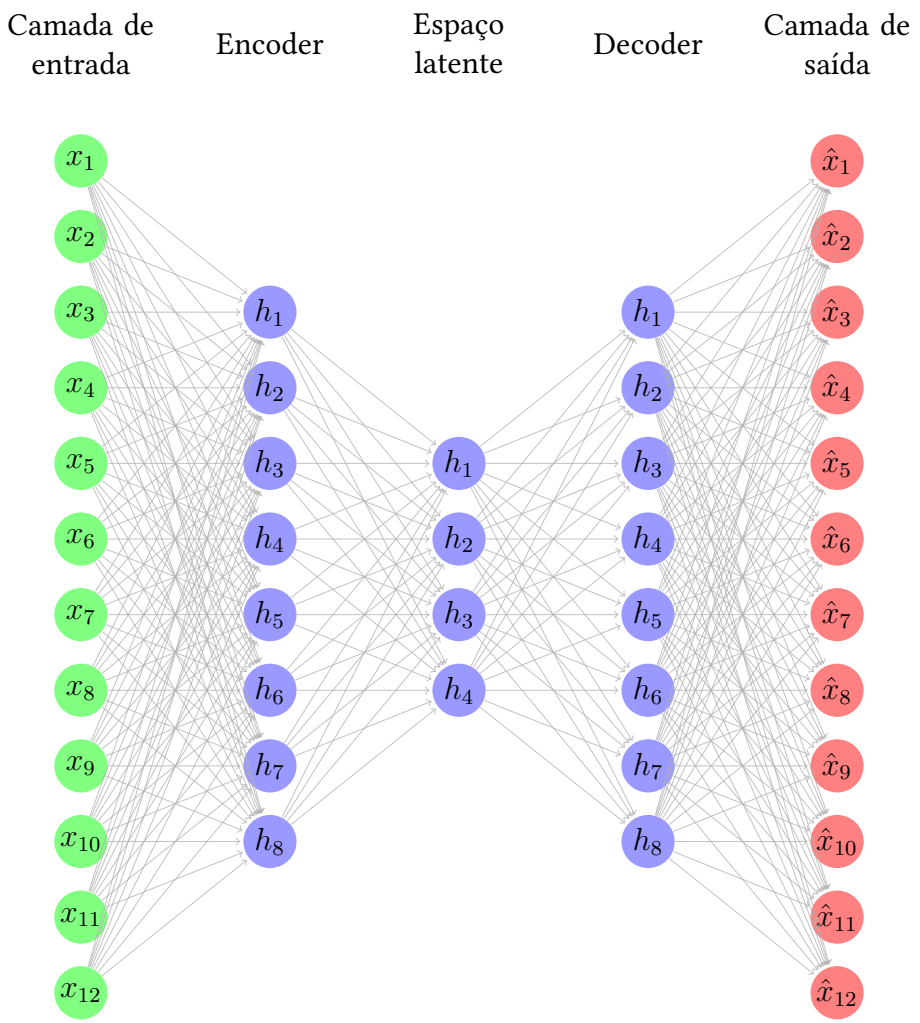

Figura 4.9: Arquitetura do autoencoder.

Capítulo 5

Resultados e Discussão

5.1 Resultados

O erro quadrático médio (MSE), quantifica a discrepância entre os dados de entrada padronizados e suas reconstruções. Para cada um dos dias utilizados no treinamento, calculou-se um MSE.

O MSE médio do modelo final, foi de aproximadamente 0,16, o que indica que, em média, o erro de reconstrução é inferior a meio desvio padrão, considerando que os dados estavam escalados.

Percentil	MSE	RMSE (σ)
50	0.03	0.16
80	0.18	0.42
90	0.34	0.58
95	0.56	0.75
99	1.59	1.26

Tabela 5.1: Erro de reconstrução do autoencoder por percentil no cluster de referência.

Ao se observar os percentis de erro na tabela 5.1, pode-se ver que, ao menos, 80% dos dias do cluster nº0, estão com erro menor que meio desvio padrão. 95% não chegam a 0,75σ.

Apesar da estrutura simples dos autoencoders, acredita-se então que foram suficientes para capturar a essência desses comportamentos, ao mesmo tempo em que se manteve abertura para comportamentos atípicos, pois pessoas do cluster nº0 também tinham dias atípicos.

Buscou-se também analisar, na tabela 5.2, dentre os usuários do cluster nº0, quais são aqueles com maior erro.

UserID	MSE Mean	MSE Std	Total days
1455047	9.33	4.43	2
1343144	8.49	21.70	22
794521	6.90	1.64	3
1249661	6.49	16.56	7
1357121	6.21	6.94	20

Tabela 5.2: Top 5 usuários do cluster nº0 com maior erro médio.

Nas figuras 5.1 e 5.2, analisou-se alguns desses usuários individualmente para entender os motivos de erros altos, mesmo estando no cluster nº0. Priorizou-se os que tinham um número maior de dias, a fim de facilitar a análise.

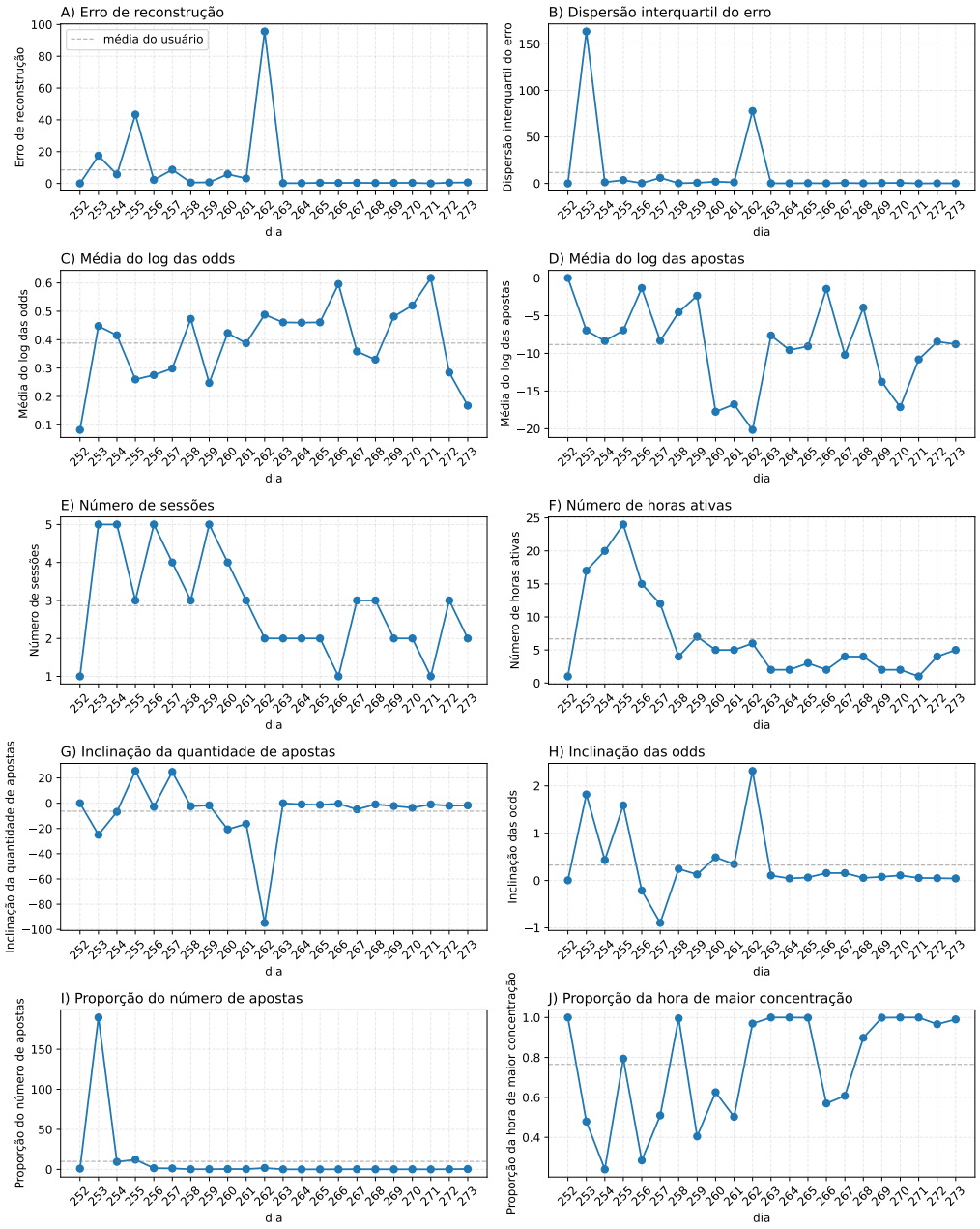

Figura 5.1: Histórico de apostas - Usuário 1343144.

No caso do usuário 1343144, ao analisar a figura 5.1, é possível observar pelo menos três dias com um erro muito acima da média: 253, 255 e 262 (A).

E nesse mesmo gráfico, é possível justificar esses erros:

- Dia 253, há um total de cinco sessões (E), um número alto de horas ativas (F), uma alta tendência de odds (H) e um número altíssimo de apostas (I);
- Dia 255, há um número altíssimo de horas ativas (F), uma alta tendência na quantidade de apostas (G) e uma alta tendência de odds (H);

- Dia 262, há uma baixíssima média do log da proporção de aposta (D) e uma baixa tendência na quantidade de apostas (G) e uma alta tendência de odds (H);

Pode-se também observar o gráfico (B) na figura 5.1, que mostra a dispersão interquartil (IQR) dos erros calculados. Nota-se principalmente que a maior divergência se concentra justamente nos dias 253 e 262, anteriormente discutidos. O dia 255 por sua vez apresenta um IQR menor em relação aos outros, demonstrando uma maior certeza que é um dia atípico.

Não se pode associar um erro alto a um comportamento problemático. A essência do modelo, da forma como foi treinado, é destacar dias atípicos. Pode-se corroborar isso, ao se observar o dia 262, que teve um erro alto, porém ao se analisar suas características, entendeu-se que se tratava apenas de um dia muito diferente dos demais, ao contrario dos dias 253 e 255, que apresentaram um comportamento mais problemático.

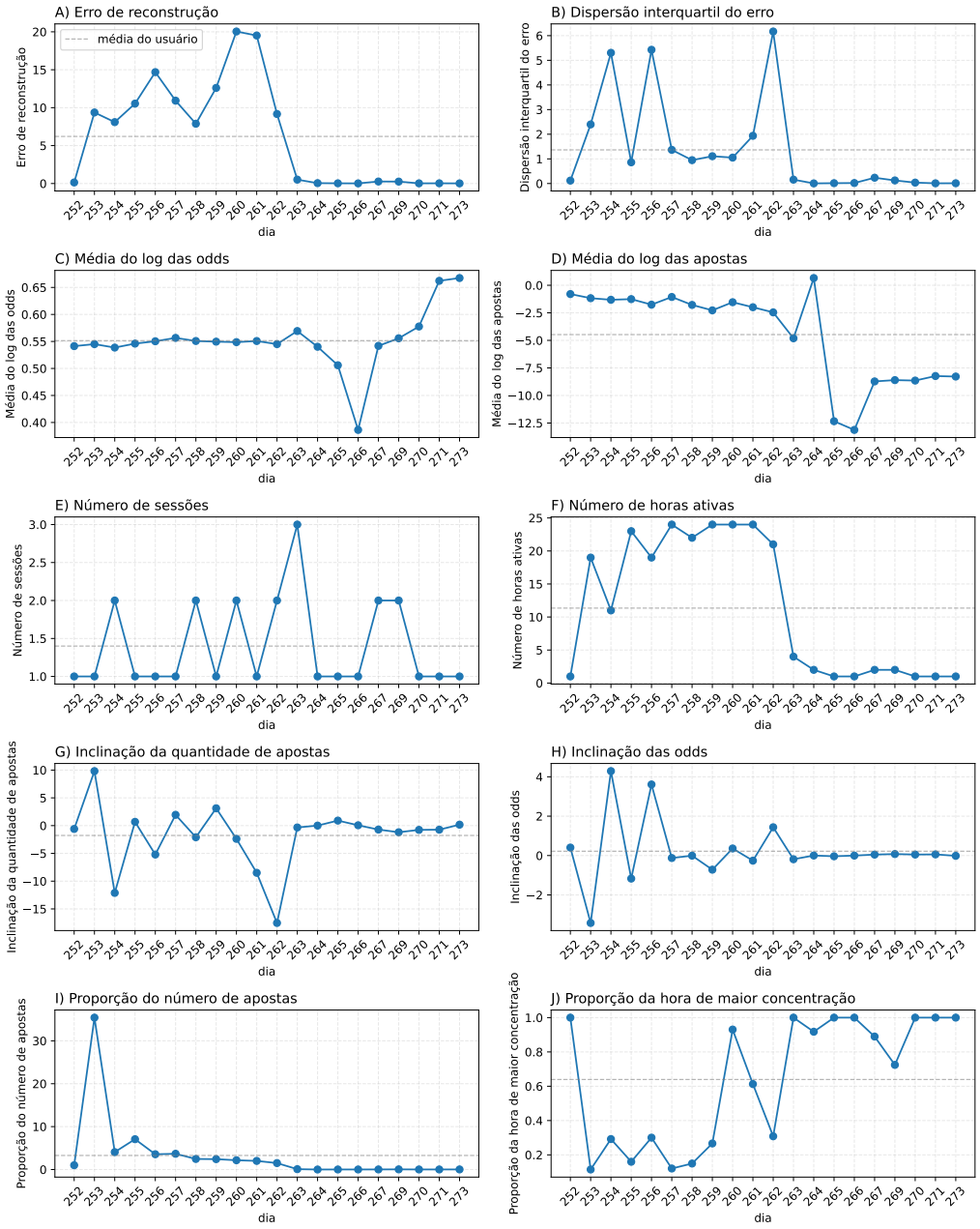

Figura 5.2: Histórico de apostas - Usuário 1357121.

Ao analisar o usuário 1357121 na figura 5.2, nota-se que ele possui dois momentos distintos de tipo de jogo: na primeira metade da amostra, ele tinha um perfil muito mais intenso e dinâmico; já na segunda metade, ele apresentava um perfil muito mais conservador e constante.

Pode-se observar que tanto o MSE (A) quanto o IQR (B), conseguiram acompanhar essas características, moldando com sucesso a mudança de comportamento do indivíduo.

Na tabela 5.3, buscou-se analisar os usuários com comportamento um pouco mais normal, com erro mais próximo do percentil 80 do cluster nº0, ou seja, 0,19.

UserID	MSE Mean	Total days
1432639	0.190598	2
1495917	0.190660	2
1234049	0.190508	6
1319390	0.190503	4
1046423	0.190480	12

Tabela 5.3: Top 5 usuários do cluster nº0 com menor distancia de erro P80.

A partir da tabela 5.3, foi escolhido o usuário 1046423 também pelo número de dias. Como se observa na figura 5.3, apesar do baixo erro geral, ainda é possível notar as nuances de comportamento do usuário ao longo do tempo.

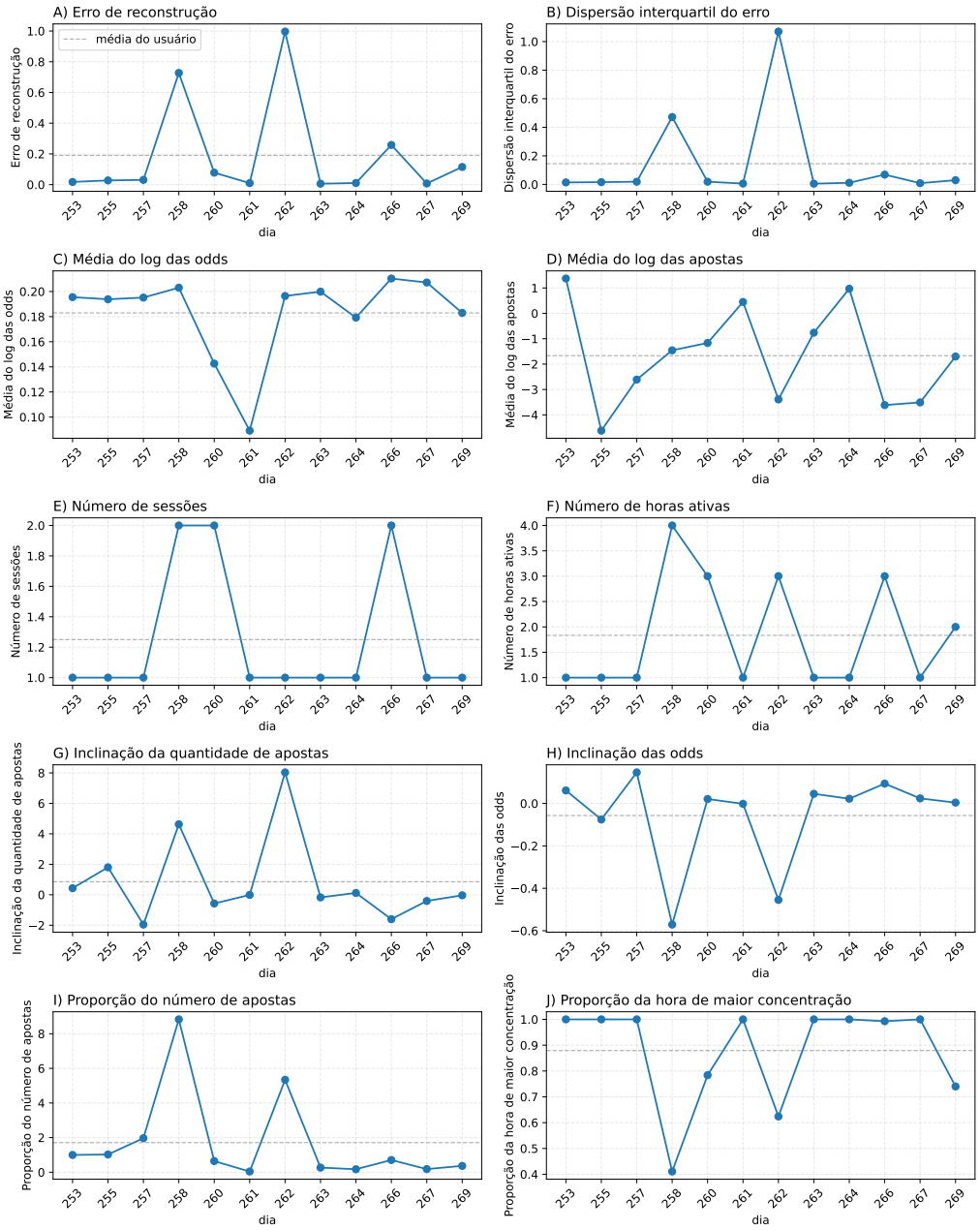

Figura 5.3: Histórico de apostas - Usuário 1046423.

5.1.1 Análise comparativa entre clusters

Após a análise do cluster nºo, passou-se para a análise de todos os clusters. Então, para cada um dos dias de aposta, calculou-se o MSE.

Cluster	Mean MSE	STD MSE	MSE P25	MSE P50	MSE P75	Max MSE
nº0	0.16	0.79	0.008	0.026	0.135	95.61
nº1	2.03	6.24	0.155	0.467	1.524	267.18
nº2	1.57	7.88	0.010	0.158	0.800	206.21
nº3	0.61	1.98	0.040	0.207	0.512	149.42
nº4	7.69	12.90	0.757	4.709	11.02	249.78
nº5	11.26	65.78	0.010	0.211	4.657	1034.17

Tabela 5.4: Distribuição do erro de reconstrução (MSE) por cluster.

Ao se analisar a tabela 5.4, observa-se primeiramente que o cluster com menor erro médio é o cluster nº0, o que era de se esperar, pois o modelo considerou apenas dias desse cluster.

O erro médio do cluster nº3 é o menor dentre os outros, o que é corroborado ao observar-se a figura 4.5, que mostra uma proximidade espacial entre os clusters nº0 e nº3.

Os erros médio dos clusters nº4 e nº5 por sua vez são os maiores, o que também é corroborado ao observar-se a figura 4.5, que mostra uma distância espacial entre os clusters nº0, nº4 e nº5.

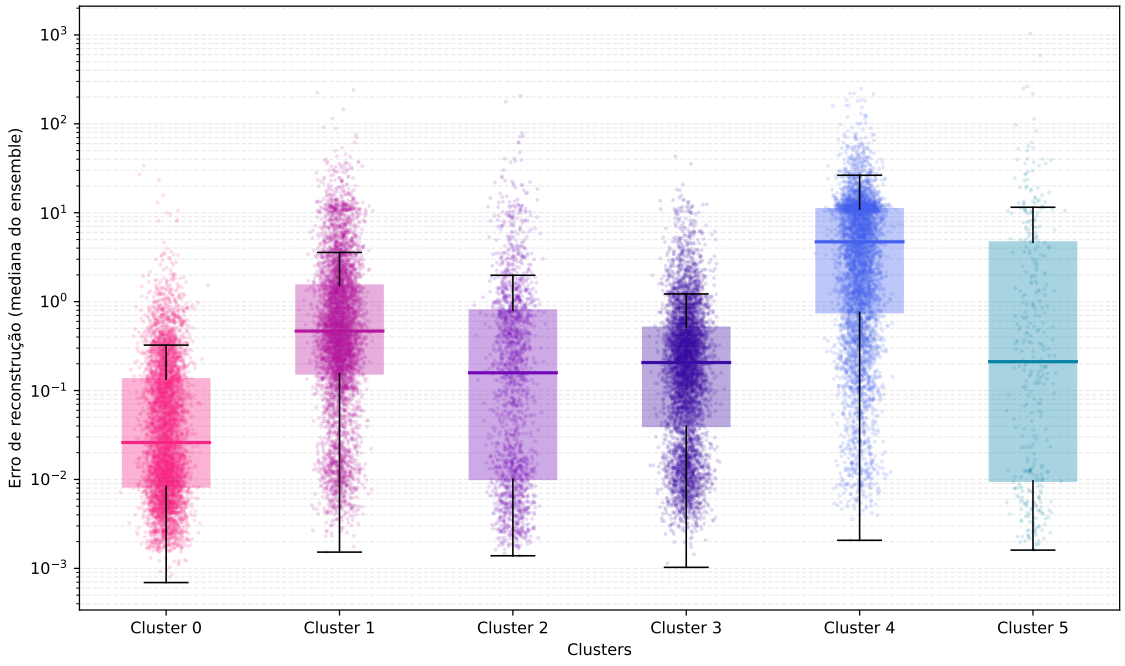

Figura 5.4: Distribuição do erro de reconstrução por cluster (escala log).

Ao se analisar a distribuição de erro dos clusters na figura 5.4, nota-se o mesmo padrão anteriormente descrito, principalmente sobre os diferentes erros em cada quartil

de cada cluster, com a diferença de que o cluster nº5 se mostra muito mais esparso que o cluster nº4, possuindo diferentes tipos de dia.

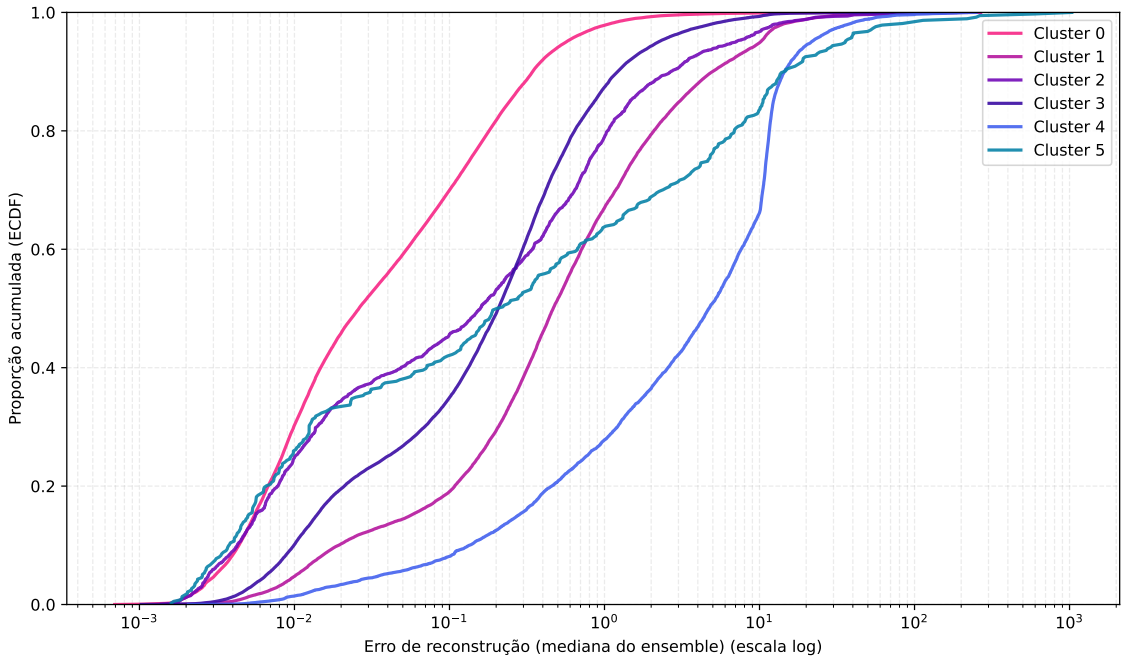

Figura 5.5: Distribuição acumulada do erro de reconstrução (ECDF) por cluster.

Ao se observar a figura 5.5, novamente destaca-se a nítida diferença de dispersão de erro dos clusters. Nota-se que alguns têm um comportamento similar em determinados momentos, enquanto certos pontos mudam bruscamente, como o caso do cluster nº5, que apesar de possuir elementos com um erro muito alto, boa parte de seus dias têm um valor normal, similar aos clusters nº0 e nº2.

O cluster nº4 de longe se mostra ser o cluster mais problematico, com a sua curva se destacando, e mostrando um alto erro, na maioria dos dias.

5.1.2 Limiar de anormalidade

Para se considerar um dia anômalo, pode-se adotar um limiar de anormalidade. Neste caso, adotou-se o valor do percentil 95 do cluster nº0, ou seja, 0,565. Dessa maneira, garantiu-se que o teto do cluster nº0 seja utilizado como o máximo aceitável para um dia normal, desconsiderando uma possível parcela anômala.

Na tabela 5.5, destacaram-se os usuários que possuíam uma proporção de 20% de dias anômalos, utilizando-se como referência o novo limiar, ordenados pelo erro médio. Pode-se então acompanhar ou entender seus comportamentos como na figura 5.6, que mostra o histórico do usuário 1288121.

User ID	Total days	Atypical days	Mean MSE	Mean IQR	Cluster
1288121	20	4	2.74	0.61	4
1032342	20	4	2.45	1.21	0
1155938	15	3	1.47	0.97	3
339431	15	3	1.45	2.30	0
723086	15	3	0.69	0.26	0

Tabela 5.5: Usuários com aproximadamente 20% de dias atípicos.

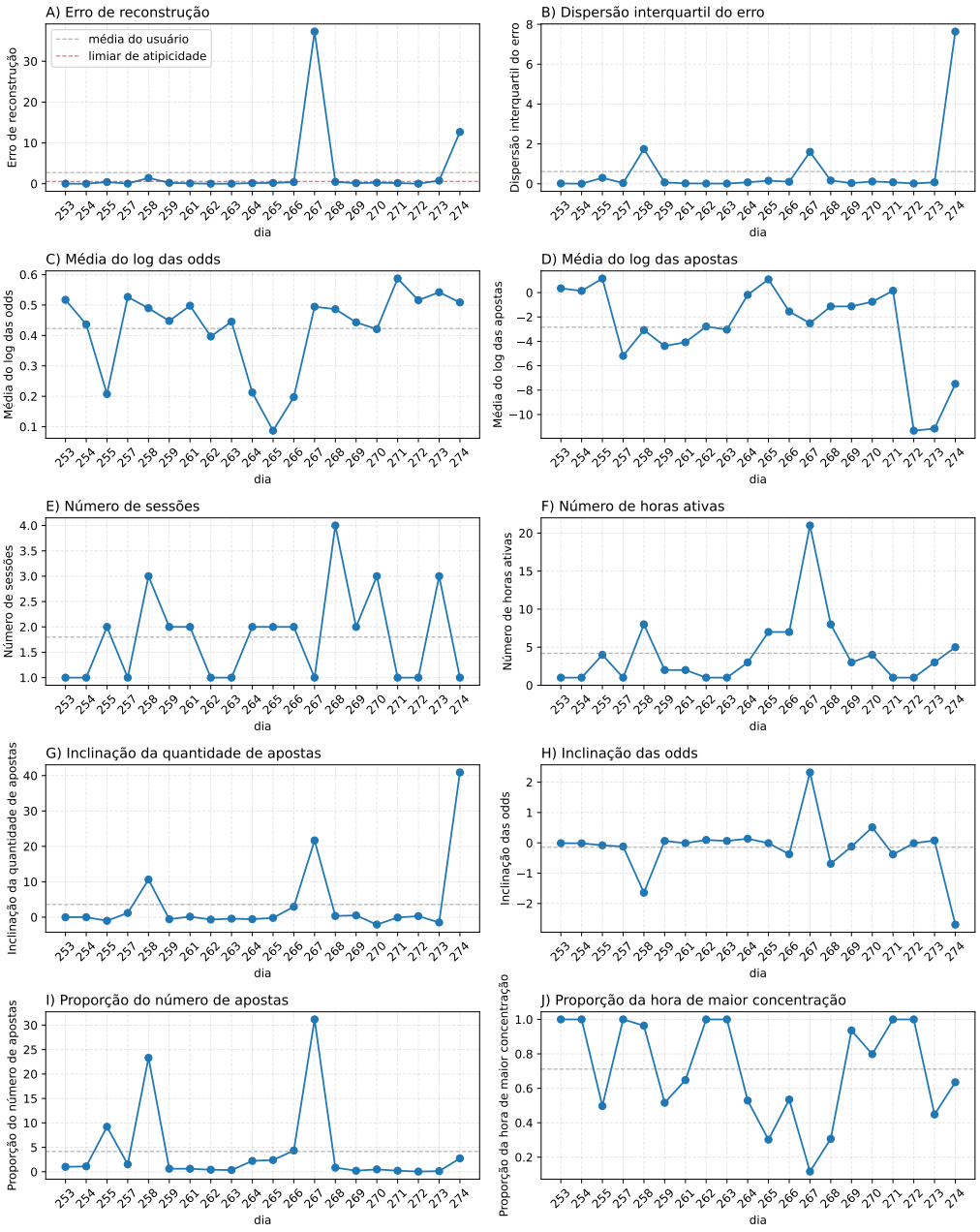

Figura 5.6: Histórico de apostas - Usuário 1288121.

User ID	Total days	Atypical days	Mean MSE	Mean IQR	Cluster
220550	18	9	11.65	5.58	1
1309287	20	10	11.55	253.37	2
1387744	6	3	11.39	4.98	1
1380472	6	3	10.75	6.98	1
1441090	6	3	10.04	6.72	4

Tabela 5.6: Usuários com aproximadamente 50% de dias atípicos.

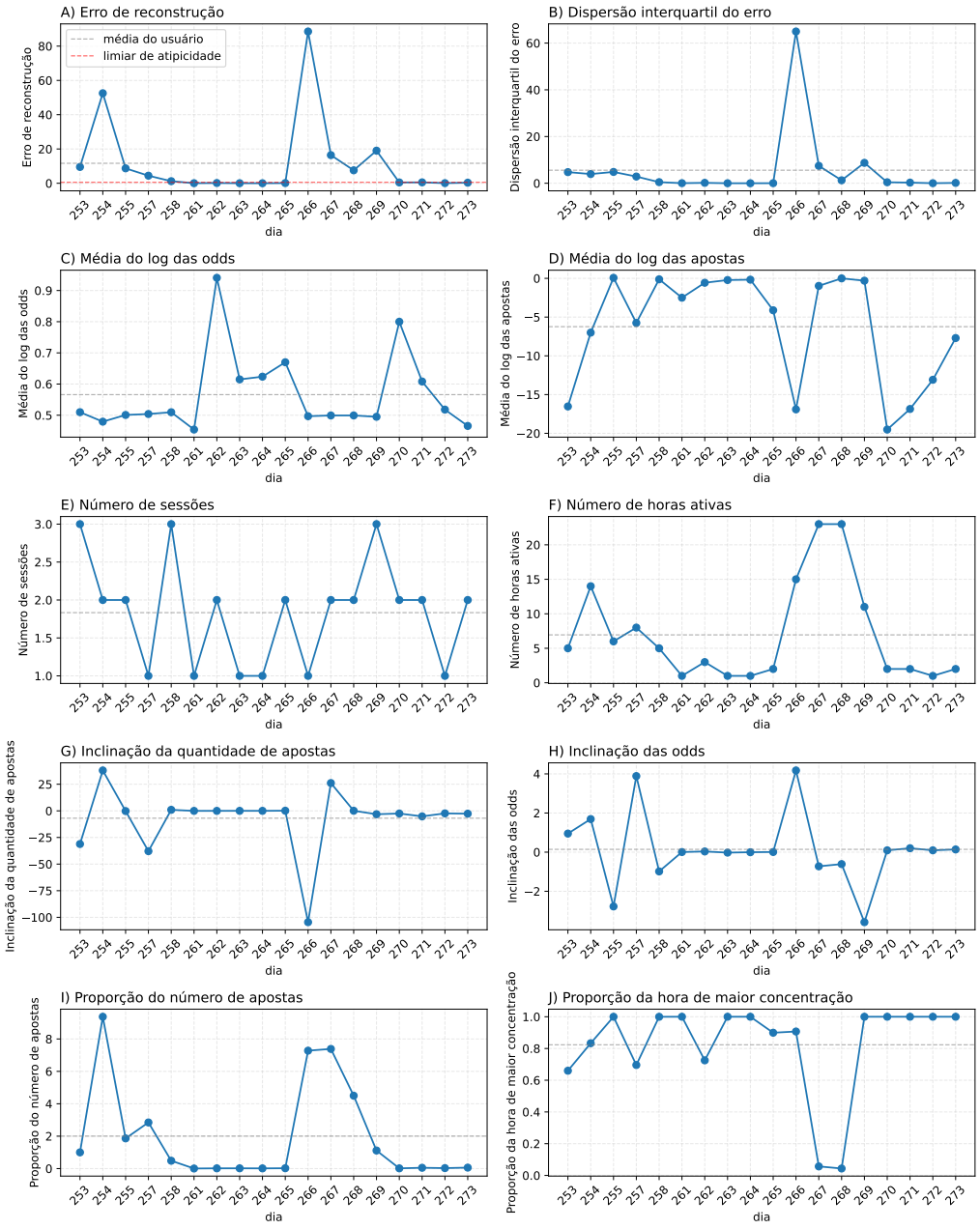

Figura 5.7: Histórico de apostas - Usuário 220550.

Na tabela 5.6, mostra-se os usuários com uma proporção de 50% de dias anômalos, também ordenados pelo erro. Destacado o usuário 220550 na figura 5.7.

Ao analisar as figuras 5.6 e 5.7, observa-se novamente que os dias com maior erro são aqueles que apresentam um comportamento mais anômalo quando comparados aos demais. O uso de um limiar também se mostra uma abordagem eficaz tanto para identificar se um dia é ou não anômalo quanto para avaliar se um indivíduo possui um histórico mais anômalo. Além disso, esse método permite verificar a ocorrência de sequências ou do aumento de dias anômalos ao longo do tempo.

5.2 Limitações

O trabalho teve como objetivo analisar o comportamento independentemente do estágio do transtorno de jogo. Portanto, não foi possível utilizar modelos supervisionados, pois, conforme discutido anteriormente, o diagnóstico dependia de o indivíduo estar em um estágio final. Isso trouxe um problema: não havia uma forma concreta de se mensurar ou comparar os resultados, dependendo-se exclusivamente da análise manual realizada nos clusters.

Outra limitação significativa diz respeito aos dados. Não existiam informações sobre o comportamento do usuário em relação a depósitos, solicitações de exclusão, solicitações de aumento de limite, entre outros aspectos. Também não se dispunha de informações demográficas, como idade e gênero, apenas informações relacionadas às apostas.

Os dados analisados correspondiam a um recorte de apenas 22 dias de histórico. Portanto, não se tinha acesso ao histórico completo de um indivíduo, não se sabia se ele era um usuário novo ou não. Além disso, não havia como garantir que, dentro dessa janela de 22 dias, houvesse indivíduos que haviam mudado de comportamento ou que esses indivíduos representassem uma parcela significativa da amostra, ainda que, segundo a literatura, esse período fosse relativamente suficiente.

Ademais, havia o problema do fuso horário: como o site contava com usuários do mundo todo, não foi possível analisar corretamente as apostas de um dia para o outro. Não era possível determinar se o usuário apostava durante o dia ou à noite, nem em quais dias da semana.

Também foi identificado o uso de apostas automáticas por parte dos usuários, o que justificava sessões com duração superior a 24 horas, embora isso não configurasse necessariamente um problema. Ainda assim, seria interessante dispor de uma variável que indicasse se uma aposta havia sido realizada de forma automática ou não.

O ideal seria que todos os dias problemáticos fossem anômalos, e que nem todos os dias anômalos fossem necessariamente problemáticos. Embora os resultados sugiram

esse comportamento, ainda há a necessidade de se validar a metodologia proposta neste trabalho.

Além disso, seria interessante investigar se apenas o erro é suficiente para identificar e acompanhar comportamentos problemáticos, ou se o uso de um limiar, bem como a escolha de seu valor, seria mais apropriado para essa finalidade.

Capítulo 6

Conclusão

Para identificar quando um apostador está mudando ou intensificando seu comportamento, buscou-se uma forma de mensurar o quanto alguém está apresentando um comportamento anormal.

Como não é possível diagnosticar todos os indivíduos, foram buscados comportamentos característicos e específicos de apostadores problemáticos, como a necessidade de intensificar sua relação com as apostas, seja aumentando o valor apostado, o número de apostas realizadas ou o tempo diário dedicado à atividade.

Ao se agrupar os jogadores com base nessas métricas de comportamento, conseguiu-se identificar com sucesso um grupo de controle, não necessariamente de jogadores que apostam pouco, mas que mantiveram seu comportamento estável ao longo do tempo.

Com a construção de um conjunto de autoencoders, capturou-se com sucesso a essência do comportamento do grupo de controle, permitindo, dado um dia de aposta, obter a distância, ou diferença do que seria normal para o grupo de controle.

Um próximo passo para a aplicação dessa técnica seria utilizar essa distância para identificar um comportamento problemático, acompanhar a distância de um usuário ao longo do tempo e, se necessário, aplicar medidas apropriadas.

A solução não se restringe ao jogo escolhido, para contemplar apostas esportivas ou outros jogos de cassino com quota fixa, basta ter formas de quantificar comportamentos problemáticos, como o risco corrido dentro de cada aposta. Nesse caso, a quantificação foi feita pela combinação da cotação e do valor apostado.

Esse processo também poderia ser replicado em outras adicções comportamentais, desde que existam formas de se analisar e quantificar o histórico de comportamento de um indivíduo, bem como acompanhá-lo ao longo do tempo.

Seria interessante aplicar esses métodos a um conjunto de dados e acompanhar os resultados ao longo do tempo, a fim de se avaliar adequadamente os grupos, os modelos e as diferentes formas de utilização dessa distância.

Referências

ANDERSSON, Sam *et al.* Insights into the temporal dynamics of identifying problem gambling on an online casino: A machine learning study on routinely collected individual account data. **Journal of Behavioral Addictions**, v. 14, n. 1, p. 490–500, 2025. ISSN 2062-5871, 2063-5303. DOI: 10.1556/2006.2025.00013. Available from: https://akjournals.com/view/journals/2006/14/1/article-p490.xml. Visited on: 25 Oct. 2025.

AUER, Michael; GRIFFITHS, Mark D. Using artificial intelligence algorithms to predict self-reported problem gambling with account-based player data in an online casino setting. **Journal of Gambling Studies**, v. 39, n. 3, p. 1273–1294, 2022. ISSN 1573-3602. DOI: 10.1007/s10899-022-10139-1. Available from: https://link.springer.com/10.1007/s10899-022-10139-1. Visited on: 24 Oct. 2025.

CHALLET-BOUJU, Gaëlle *et al.* Modeling Early Gambling Behavior Using Indicators from Online Lottery Gambling Tracking Data: Longitudinal Analysis. **Journal of Medical Internet Research**, v. 22, n. 8, e17675, 2020. ISSN 1438-8871. DOI: 10.2196/17675. Available from: http://www.jmir.org/2020/8/e17675/. Visited on: 25 Oct. 2025.

GHELFI, Michela *et al.* Online Gambling: A Systematic Review of Risk and Protective Factors in the Adult Population. **Journal of Gambling Studies**, v. 40, p. 673–699, 2024. ISSN 1573-3602. DOI: 10.1007/s10899-023-10258-3. Available from: https://link.springer.com/10.1007/s10899-023-10258-3. Visited on: 3 Dec. 2025.

GRANT, Jon E. *et al.* Introduction to Behavioral Addictions. **The American Journal of Drug and Alcohol Abuse**, v. 36, n. 5, p. 233–241, 2010. DOI: 10.3109/00952990.2010.491884.

GÜELL, Francisco. The liberating dimension of human habit in addiction context. **Frontiers in Human Neuroscience**, v. 8, 2014. ISSN 16625161. DOI: 10.3389/fnhum.2014.00664. Available from: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00664/abstract. Visited on: 4 Dec. 2025.

KOOB, George F.; LE MOAL, M. Drug Addiction, Dysregulation of Reward, and Allostasis. **Neuropsychopharmacology**, v. 24, n. 2, p. 97–129, 2001. ISSN 0893133X. DOI: 10.1016/S0893-133X(00)00195-0. Available from: https://www.nature.com/doifinder/10.1016/S0893-133X(00)00195-0. Visited on: 11 Dec. 2025.

LAJČINOVÁ, Bibiána; GALL, Marián; MICHAL, Pitoňák. **Anomaly Detection in Time Series Data: Gambling prevention using Deep Learning**. [*S. l.*], 2023. Available from: https://eurocc.nscc.sk/wp-content/uploads/2023/08/nscc_gambling_en.pdf. Visited on: 18 Dec. 2025.

LAMB, R.J.; GINSBURG, Brett C. Addiction as a BAD, a Behavioral Allocation Disorder. **Pharmacology Biochemistry and Behavior**, v. 164, p. 62–70, 2018. ISSN 00913057. DOI: 10.1016/j.pbb.2017.05.002. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0091305717300655. Visited on: 3 Dec. 2025.

PERES, Fernando *et al.* Time Series Clustering of Online Gambling Activities for Addicted Users' Detection. **Applied Sciences**, v. 11, n. 5, p. 2397, 2021. ISSN 2076-3417. DOI: 10.3390/app11052397. Available from: https://www.mdpi.com/2076-3417/11/5/2397. Visited on: 25 Oct. 2025.

SANTOS ZAVA, Daiane Marcele Rêis dos; SOUZA, Beatriz de Barros; MESSETTI, Paulo André Stein. The phenomenon of electronic betting games: socioeconomic and health impacts - an integrative literature review. **Clinics Biopsychosocial**, v. 2, n. 2, p. 149–154, 2024. ISSN 2965-5986. DOI: 10.54727/cbps.2.2.53. Available from: https://clinicsbiopsychosocial.com/index.php/01/article/view/39. Visited on: 24 Oct. 2025.

SOLLY, Jeremy E. *et al.* Recent advances in understanding how compulsivity is related to behavioural addictions over their timecourse. **Current Addiction Reports**, v. 12, n. 1, p. 26, 2025. ISSN 2196-2952. DOI: 10.1007/s40429-025-00621-2. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11850568/. Visited on: 19 Oct. 2025.

SUZUKI, Hiroko *et al.* Early Detection of Problem Gambling based on Behavioral Changes using Shapelets. *In:* PROCEEDINGS of the IEEE/WIC/ACM International

Conference on Web Intelligence (WI 2019). Thessaloniki, Greece: ACM, 2019. p. 367–372. DOI: 10.1145/3350546.3352549. Available from: https://dl.acm.org/doi/10.1145/3350546.3352549. Visited on: 25 Oct. 2025.

THAKUR, Pratibha; KASHYAP, Somendra Singh. Behavioral Addiction: A Review of Current Understanding and Emerging Perspectives. **Zeichen Journal**, v. 11, n. 1, p. 1–16, 2025. Available from: https://www.researchgate.net/publication/388066147_Behavioral_Addiction_A_Review_of_Current_Understanding_and_Emerging_Perspectives.