Identifying Gambling Risk Profiles from Online Behavioral Data: An AI Cluster-Based Empirical Study

Arthur C. S. Xavier[1], **Milton T. M. Junior**[1], **Hendrik W. C. Garcia** [3], **Adriana F. G. Barreto**[1], **Tiago A. E. Ferreira**[2]

[1]Futuro Tech – Recife, Brazil

[2]Department of Statistics and Informatics (DEINFO)
Federal Rural University of Pernambuco – Recife, Brazil

[3]Department of Psychology
Pernambuco Faculty of Health (FPS) – Recife, Brazil

{arthur.xavier,adriana.falcao}@futuroai.tech, tiago.espinola@ufrpe.br

{hendrikgarcia97,drmiltonpsiquiatra}@gmail.com

Abstract. *This study investigates whether behavioral risk profiles can be identified using clustering techniques applied to real-world player tracking data. Only in 2023, more than 7% of the Brazilian population was classified as high risk for problematic gambling. Behavioral indicators were extracted from the transactional data of more than 11,000 gamblers from a Brazilian online operator and used to train a K-Means clustering model. The resulting clusters were evaluated using a daily monitoring dataset and compared with a group of voluntary self-excluded players. The results show that approximately 89% of these players were classified in the high-risk clusters at least once, indicating that the proposed indicators capture behavioral patterns associated with gambling risk.*

1. Introduction

Gambling is defined as risking money — or other items of value — on an event of uncertain outcome, with the possibility of gaining an increased return [American Psychiatric Association 2022]. In Brazil, with transactions exceeding US$3.4 billion per month in 2024 [Banco Central do Brasil 2024], the betting market has positioned the country among the largest in the world and continues to expand. Gambling Disorder, defined by the DSM-5-TR as a persistent pattern of gambling behavior causing clinically significant distress or impairment [American Psychiatric Association 2022], already affects approximately 10.9 million Brazilians (7.3% of the population), with more than 50% of these bettors having monthly incomes below US$231.38 [Universidade Federal de São Paulo (UNIFESP) 2025].

Given the seriousness of the issue and the high impact it has on society, novel approaches for early identification of problematic gambling can help mitigate the losses of online bettors. As a response to the observed challenges, this study aims to investigate whether it is possible to group gamblers into different risk profiles based on real data from an online Brazilian casino operator with more than 11,000 gamblers. Unlike previous studies focusing on isolated indicators or direct approximations of clinical criteria, this work integrates multiple behavioral proxies into a unified feature framework and applies clustering techniques to identify risk-oriented behavioral profiles.

This paper is structured as follows. Section 2 discusses the different works to which this study is related, highlighting the differences and objectives of each. Section 3 presents the methodology used to carry out the tests. Section 4 outlines the experiments with the new features and clustering models. Finally, Section 5 summarizes the study and discusses potential directions for future research.

2. Related Works

In a study conducted by [Catania and Griffiths 2021], the authors investigated whether DSM-5-TR criteria for Gambling Disorder can be approximated using online gambling tracking data. Using behavioral indicators derived from transactional records, the authors applied a two-step cluster analysis to identify four behavioral gambler profiles. The study highlights the difficulty of operationalizing several clinical criteria using purely behavioral data. While this work demonstrates the potential of behavioral tracking data for identifying gambling risk patterns, the difficulty in operationalizing several DSM-5-TR criteria highlights the challenge of translating clinical constructs into observable behavioral indicators. A related line of research focuses on defining specific behavioral proxies for key harm mechanisms.

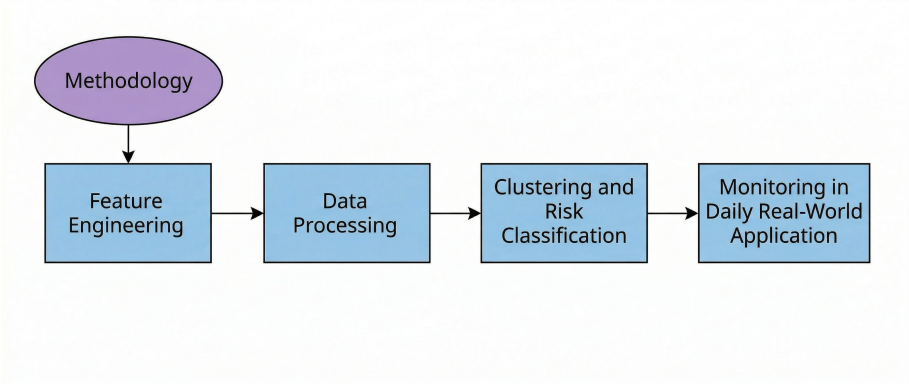
The study of [Auer and Griffiths 2023] aimed to operationalize the behavioral indicator of loss chasing, one of the key harm markers in gambling disorder. Using account-based tracking data from an online casino operator, the authors analyzed behavioral records, including bets, sessions, and monetary deposits. The study proposed several behavioral operationalizations of loss chasing and evaluated them using statistical tests and regression models relating the indicators to behavioral risk categories. The results suggested that frequent session depositing was the most consistent behavioral proxy associated with loss chasing. However, the study evaluated the metrics individually rather than integrating them into a broader behavioral risk framework.

Building on these studies, the present work proposes a structured set of behavioral indicators capturing gambling engagement, intensity, and chasing dynamics. Loss chasing is modeled using two behavioral indicators, while a regression-based trend metric approximates the DSM-5-TR tolerance criterion. Additional features capture session patterns, betting intensity, and temporal engagement. These indicators are used in a K-Means clustering framework to identify behavioral risk profiles and evaluate them in a real-world monitoring scenario. The results show that more than 80% of voluntary self-excluded players are concentrated in the highest-risk clusters.

3. Methodology

This study adopts a quantitative approach for gambling risk classification using an anonymized dataset from a Brazilian gambling operator containing data from October 2024 to November 2025. The methodology follows four main steps (Figure 1). First, feature engineering was applied to extract behavioral indicators capturing gambling volume, loss-chasing dynamics, and temporal engagement. Next, the dataset was filtered to retain only relevant observations and standardized using z-score normalization to ensure comparability between indicators.

Figure 1. Methodology Diagram



After preprocessing, the data was submitted to K-Means clustering, a centroid-based algorithm used to group gamblers according to behavioral patterns observed in transactional data. Data from October 2024 to August 2025 was used for model training, while the remaining period was used for evaluation in a daily monitoring scenario. Cluster centroids were used to estimate relative risk levels, and the model was evaluated by analyzing the distribution of voluntary self-excluded (VSE) players across the identified clusters.

Table 1. K-Means Training Hyperparameters

Hyperparameter	Value
Number of Clusters (K)	5
Random State	42
Number of Initializations (n_init)	auto
Use KMeans++	True

4. Results and Discussion

4.1. Feature Engineering

Session-based features are computed by defining a gambling session as a sequence of bets separated by no more than 30 minutes of inactivity, a threshold commonly used to approximate continuous gambling activity. Session-level statistics are then aggregated to construct player-level behavioral indicators guided by DSM-5-TR proxies, with tolerance approximated through the trend of bet values over time and chasing behavior modeled through deposit frequency within sessions.

The tolerance criterion was operationalized by fitting a linear regression to each gambler's stake values over time, using the slope coefficient (β) as an indicator of progressive increases, where $\beta_i > 0$ suggests a rising betting pattern.

Following [Auer and Griffiths 2023], loss chasing was operationalized through two frequency-based indicators — session deposit frequency and loss chasing frequency (deposits after a loss). A third indicator, loss chasing intensity, refers to the increase in stake size after a loss, approximating the behavioral escalation pattern associated with chasing dynamics.

Although several DSM-5-TR criteria cannot be directly operationalized using the available transactional data, additional behavioral features were included to capture different dimensions of gambling engagement. Table 2 summarizes the 12 behavioral indicators used in this work. Together, these indicators help characterize the intensity, temporal patterns, and behavioral variability of gamblers interacting with the operator's platform.

Table 2. Behavioral Indicators Used in the Study

Indicator	Description
avg_total_sessions	Average number of sessions per active day
avg_session_length	Mean duration of gambling sessions
std_session_length	Variability of session duration
deposits_per_session	Average number of deposits within a session
loss_chasing_frequency	Frequency of deposits after losses
loss_chasing_intensity	Stake increase after losses
pct_nightly_play	Proportion of bets placed at night
coefficient_of_variation_stake	Variability of bet values
game_types	Number of different gambling game types played
proportion_session_amount	Ratio between total bet and deposited values
bet_value_trend_beta	Trend coefficient of stake value over time
bet_value_trend_r2	Determination coefficient of the trend model

4.2. Data Processing

After feature engineering, preprocessing steps were applied to improve clustering quality. Only players active for at least three days were included in the training dataset. Additionally, the top 1% of extreme outliers were removed to reduce distortion in indicators highly sensitive to extreme values. After filtering, the dataset was reduced from approximately 35,000 to 11,641 players. Finally, features were standardized using StandardScaler.

4.3. Clustering and Risk Classification

With the data ready, the clustering algorithm was applied. After training the model, 5 clusters were generated. Cluster risk levels were estimated using the mean value of each cluster centroid, assuming that higher feature values correspond to higher behavioral risk. Also, it is important to note that it is assumed each feature has the same significant impact on the risk level. With the ranking, cluster 0 represented 52.06% of the training database, cluster 1 16.81%, cluster 2 22.65%, cluster 3 7.46%, and cluster 4 1.02%

Cluster centroids reveal clear behavioral differences between risk groups. Higher-risk clusters show greater engagement, more frequent sessions, stronger loss-chasing patterns, and higher bet-to-deposit ratios. Lower-risk clusters present lower engagement, fewer sessions, and minimal chasing behavior. One curious fact observed in the clustering of the lower cluster is that the algorithm grouped gamblers with less engagement but allowed some of the gamblers to have a higher loss-chasing intensity (i.e., stake rise after losses); however, since their other indicators are lower, it is assumed that having a stake rise after a loss with minimal engagement on the platform may indicate isolated stake increases rather than persistent chasing behavior.

The regression-based trend indicator produced coefficients close to zero across clusters. This result suggests that bet-level granularity limits the ability of linear regression to capture meaningful betting escalation patterns. Aggregating betting activity over longer periods (e.g., weekly or monthly) may yield more expressive trend coefficients and should be explored in future work. The cluster centroids can be viewed in Table 3.

Table 3. Cluster Centroids

	Cluster rank				
Feature	**0**	**1**	**2**	**3**	**4**
game_type	0.38	0.34	0.67	3.60	1.16
avg_total_sessions	1.34	1.28	2.33	1.87	2.86
avg_session_length	0.21	0.22	0.42	0.88	0.57
std_session_length	0.21	0.18	0.48	0.93	0.65
pct_nightly_play	0.00	0.07	0.00	0.00	0.00
deposits_per_session	0.08	0.08	0.17	0.44	2.67
loss_chasing_frequency	0.12	0.11	0.35	0.68	5.46
loss_chasing_intensity	12.80	5.14	3.16	10.44	6.83
coefficient_of_variation_stake	0.01	0.02	0.00	0.00	0.00
proportion_session_amount	3.38	3.45	8.00	6.97	5.41
bet_value_trend_beta	<0.01	0.00	<0.01	0.00	0.00
bet_value_trend_r2	0.12	0.73	0.14	0.22	0.26

4.4. Monitoring in Daily Real-World Application

After the clustering, the model was used to classify different players in a daily monitoring dataset. The test dataset represented 11,883 different gamblers with more than 84,000 daily observations, each observation representing an active day for a gambler. To test the potential effectiveness of the model, it was granted access to a group of players who were voluntarily self-excluded. This group was composed of a total of 322 gamblers, but in the monitoring dataset, it was able to access 132 gamblers with 789 observations. After applying the model to the monitoring dataset, it was observed that approximately 89% of the VSE group was classified in the higher risk cluster at least once, indicating that the model could possibly signal and notify problematic behavior before the gambler had to take more serious measures or even reach severe levels of harm (Table 4).

5. Final Considerations

This study investigated whether clustering techniques applied to behavioral indicators can identify gambling risk profiles using real-world transactional data. The results indicate that the proposed indicators capture meaningful behavioral patterns associated with gambling risk, as evidenced by the concentration of voluntary self-excluded players in higher-risk clusters.

Although the results are promising, alternative clustering methods and distance metrics should be explored in future work. Moreover, even though the use of VSE groups is a common proxy for gambling harm [Bijker et al. 2023], voluntary self-exclusion may

occur for reasons other than gambling disorder, suggesting that additional ground-truth indicators should be investigated.

However, the study already showed that there is a promising path for constructing more robust classification models using player-level behavioral indicators and clustering algorithms to create gambling profiles with different levels of risk, ensuring the continuity of efforts to achieve the objective of mitigating higher cases of gambling disorders in society.

Table 4. Distribution by *cluster_rank* in daily observations vs. highest level reached by players

cluster_rank	Daily observations		Players by highest level reached	
	n	%	n	%
0	12	1.52	6	4.55
1	16	2.03	2	1.52
2	41	5.20	7	5.30
3	1	0.13	0	0.00
4	719	91.13	117	88.64
Total	**789**	**100.00**	**132**	**100.00**

References

American Psychiatric Association (2022). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association Publishing, Washington, DC.

Auer, M. and Griffiths, M. D. (2023). An empirical attempt to operationalize chasing losses in gambling utilizing account-based player tracking data. *Journal of Gambling Studies*, 39:1547–1561.

Banco Central do Brasil (2024). Análise técnica sobre o mercado de apostas online no brasil e o perfil dos apostadores [technical analysis of the online betting market in brazil and the profile of bettors]. Estudo Especial 119, Banco Central do Brasil, Brasília.

Bijker, R., Booth, N., Merkouris, S. S., and et al. (2023). International prevalence of self-exclusion from gambling: a systematic review and meta-analysis. *Current Addiction Reports*, 10:844–859.

Catania, M. and Griffiths, M. (2021). Applying the dsm-5 criteria for gambling disorder to online gambling account-based tracking data: An empirical study utilizing cluster analysis. *Journal of Gambling Studies*, 38:1–18.

Universidade Federal de São Paulo (UNIFESP) (2025). *Caderno Temático LENAD III: Jogos de Aposta na População Brasileira – Resultados 2023 [LENAD III Thematic Report: Gambling in the Brazilian Population – 2023 Results]*. UNIAD/UNIFESP, São Paulo.