AKADÉMIAI KIADÓ

Development and validation of a prediction model for online gambling problems based on players' account data

Journal of Behavioral Addictions

11 (2022) 3, 874–889

DOI:
10.1556/2006.2022.00063

FULL-LENGTH REPORT

BASTIEN PERROT[1,2], JEAN-BENOIT HARDOUIN[1,2], ELSA THIABAUD[3], ANAÏS SAILLARD[3], MARIE GRALL-BRONNEC[1,3] and GAËLLE CHALLET-BOUJU[1,3]*

[1] Nantes Université, Univ Tours, CHU Nantes, CHU Tours, INSERM, MethodS in Patients Centered Outcomes and HEalth ResEarch, SPHERE, F-44000, Nantes, France

[2] Nantes Université, CHU Nantes, Biostatistics and Methodology Unit, Department of Clinical Research and Innovation, F-44000, Nantes, France

[3] Nantes Université, CHU Nantes, UIC Psychiatrie et Santé Mentale, F-44000, Nantes, France

Received: October 5, 2021 • Revised manuscript received: March 3, 2022; June 1, 2022; August 1, 2022 • Accepted: August 13, 2022
Published online: September 20, 2022

ABSTRACT

Background and aims: Gambling disorder is characterized by problematic gambling behavior that causes significant problems and distress. This study aimed to develop and validate a predictive model for screening online problem gamblers based on players' account data. *Methods:* Two random samples of French online gamblers in skill-based (poker, horse race betting and sports betting, n = 8,172) and pure chance games (scratch games and lotteries, n = 5,404) answered an online survey and gambling tracking data were retrospectively collected for the participants. The survey included age and gender, gambling habits, and the Problem Gambling Severity Index (PGSI). We used machine learning algorithms to predict the PGSI categories with gambling tracking data. We internally validated the prediction models in a leave-out sample. *Results:* When predicting gambling problems binary based on each PGSI threshold (1 for low-risk gambling, 5 for moderate-risk gambling and 8 for problem gambling), the predictive performances were good for the model for skill-based games (AUROCs from 0.72 to 0.82), but moderate for the model for pure chance games (AUROCs from 0.63 to 0.76, with wide confidence intervals) due to the lower frequency of problem gambling in this sample. When predicting the four PGSI categories altogether, performances were good for identifying extreme categories (non-problem and problem gamblers) but poorer for intermediate categories (low-risk and moderate-risk gamblers), whatever the type of game. *Conclusions:* We developed an algorithm for screening online problem gamblers, excluding online casino gamblers, that could enable the setting of prevention measures for the most vulnerable gamblers.

KEYWORDS

gambling, prediction model, machine learning, problem gambling, online gambling

In the original online version of this article the in-text citations appeared in an incorrect order. Therefore, it was replaced with a corrected version on 8 December 2022 without modifying the scientific content. For more about the nature of the correction see the linked erratum: https://doi.org/10.1556/2006.2022.10000

*Corresponding author.
Tel.:+33(0) 2 40 84 76 20.
E-mail: Gaelle.BOUJU@chu-nantes.fr

AKJournals

INTRODUCTION

Online gambling is a well-known risk factor for the development of gambling problems compared to land-based activity (Kairouz, Paradis, & Nadeau, 2012; Papineau et al., 2018). Explanations include high accessibility (a few seconds for opening a web page, available 24/24 7/7), privacy (no other gamblers present physically, possibility to gamble with a pseudonym), frequent gambling outcomes (higher number of gambling opportunities), and the use of digital money (use of an e-wallet to deposit, rather than having to get cash out of an ATM) (S. M. Gainsbury, 2015; Griffiths, 2003). In France, since the legalization of online gambling in 2010,

problem gambling rates among the population of past-year gamblers have considerably increased (J.-M. Costes, Richard, & Eroukmanoff, 2020). During the first years after legalization (2010–2014), "moderate-risk gambling" increased to reach 3.9% of current gamblers (J.-M. Costes, Eroukmanoff, Richard, & Tovar, 2015), with excessive gambling rates being stable (0.9%). In the following years (2014–2019), excessive gambling rates has almost doubled to reach 1.6%, while moderate-risk gambling rates remained relatively stable (4.4%) (J.-M. Costes et al., 2020).

At the same time, online gambling provides interesting opportunities to study actual gambling behavior by using operators' routinely collected data (Deng, Lesch, & Clark, 2019; S. Gainsbury, 2011). Although they come with some limitations, related for example to the lack of contextual factors, ethical issues (protection of participants' privacy or difficulty obtaining informed consent) or methodological problems (multiple accounts for one gambler or multiple gamblers for one account) [for a detailed analysis of advantages and disadvantages of account-based gambling data, see (S. Gainsbury, 2011)], players' account-based gambling data are considered more reliable and less biased than self-reported online gambling behaviors (Braverman, Tom, & Shaffer, 2014; Catania & Griffiths, 2021; S. Gainsbury, 2011; Heirene, Wang, & Gainsbury, 2021).

Several studies have explored players' account data to better understand online gambling behavior and/or to find indicators of problematic gambling activity. Many of the latter were reviewed by Deng et al. (2019). Other relevant studies include the works of Luquiens et al. (2016), Perrot, Hardouin, Grall-Bronnec and Challet-Bouju (2018), Ukhov, Bjurgert, Auer and Griffiths (2021), Challet-Bouju et al. (2020), and Kainulainen (2021). However, the definition of "people having gambling problems" was very broad. Indeed, problematic or excessive gambling was often defined by a behavioral proxy such as high involvement, temporary self-exclusion or self-reported gambling problems as the reason for closing an account. This approach suffers from limitations, as noted by Auer and Griffiths (2016). As adopted by Luquiens et al. for online poker players (Luquiens et al., 2016), a validated screening tool examining problems related to gambling rather than behavior, such as the Problem Gambling Severity Index (PGSI) (Ferris & Wynne, 2001), would allow the prediction of a more relevant outcome. This methodology would generate an explicit statistical framework, implying the development of a prediction model. The PGSI is a widely used problem gambling questionnaire that is considered as the benchmark to estimate prevalence rates of problem gambling in epidemiological studies, such as in France (J. M. Costes et al., 2011; J.-M. Costes et al., 2015, 2020). Indeed, compared to other measures such as the South Oaks Gambling Screen (SOGS) (Lesieur & Blume, 1987) or the National Opinion Research Center DSM-IV Screen for Gambling Problems (NODS) (Gerstein et al., 1999), the PGSI is based on an ordinal scaling and is not focused on diagnosis of gambling disorder. As a consequence, it allows identifying four levels of gambling severity (non-problem, low-risk, moderate-risk, and problem gambling) in the general population, thus covering the whole continuum of gambling problems, including subthreshold forms of problem gambling (N. V. Miller, Currie, Hodgins, & Casey, 2013). It has been well validated in several independent samples and with several statistical approaches (Ferris & Wynne, 2001; McCready & Adlaf, 2006; N. V. Miller et al., 2013) and was developed to identify gamblers at risk for developing problem gambling (N. V. Miller et al., 2013). As such, it is particularly adapted to serve as a gold standard for the prediction of gambling problems. However, the PGSI suffered from one major limit, which was the poor discriminating abilities between the two intermediate categories (low and moderate risk), especially due to the limited range of scores within these categories (Currie et al., 2013). Currie et al. demonstrated that this limit may be overcome by raising the threshold for the moderate-risk category to 5 instead of 3 (Currie et al., 2013).

Given the number of gamblers whose activity can be observed and the richness of players' account-based data, supervised machine learning algorithms appear to be an interesting option to identify individuals with gambling problems (Percy et al., 2016; Philander, 2014). Moreover, studies based on the analysis of gamblers' account data were often restricted to a single type of gambling (for example, only sports betting or only poker), using data from a single gambling operator. This is due to the difficulty for researchers to access gambling data from operators, and the virtual impossibility to link a player's account data from distinct operators. Thus, the observed gambling behavior might not reflect the complete online gambling behavior of an individual. A possible way to overcome this limit is to gather data from national regulatory authorities, when they exist. Indeed, they usually store account-based gambling data from all operators for the purposes of regulatory compliance checks, but may also send them to research teams under certain conditions when authorized by law and justified by the interest of the study.

The present study is the second phase of the EDEIN project (*Etude de Dépistage des comportements Excessifs de jeu sur Internet*; *Screening for Excessive Gambling Behaviours on the Internet*) (Perrot et al., 2017). Here, we developed and validated a prediction model for online gambling problems based on players' account data, and addressed the aforementioned limitations (i.e. using a clinical definition of problematic gambling, considering the full range of authorized online gambling activity, and using appropriate statistical methods).

METHODS

Participants and data

Origin of the datasets. The French regulation since 2010 provides that only four types of gambling are authorized online: poker, horse race betting, sports betting and lotteries

(including draws, bingo and scratch games). All other forms of online gambling, especially online casino games (online slot machines, online table games except poker), have always been banned in France. On the one hand, among the four authorized types of online gambling, only poker, horse race betting and sports betting are opened to competition, in the framework of a license-based system managed by the Regulatory Authority for Online Gambling (Autorité de Régulation des Jeux En Ligne, ARJEL). In its capacity as the national regulator, the ARJEL is authorized to collect and store account-based gambling data from all licensed operators. On the other hand, lotteries are subjected to a monopoly from the historical national operator (Française des Jeux, FDJ), that was not regulated by the ARJEL before 2020.

Given the French regulations, we requested account-based data from both ARJEL and the FDJ. The ARJEL dataset contained individual-level data from all authorized operators. The data covered poker, horse race betting, and sports betting. If a gambler had multiple accounts (e.g., across multiple operators), then the ARJEL aggregated the data across all of the accounts. The FDJ dataset contained individual-level data related only to lotteries. This approach allowed us to cover the whole range of online gambling activities authorized in France rather than being operator-specific. Moreover, this allowed us to develop the prediction model separately for gambling forms that involve skill (sports and horse race betting, poker) and for pure chance games (lotteries) (Bjerg, 2010). The architecture of those two datasets is given in Table S1 of the supplementary material.

Recruitment. The ARJEL sent an email to a random sample of 840,797 online gamblers who had an active gambling account (i.e., had placed at least one bet during the previous twelve months) in the competition market in two successive waves (November 2015 and February 2016). The e-mail contained an invitation to respond to an online survey hosted by ARJEL. A total of 9,306 gamblers (1.1% of those invited to participate) responded to the whole survey.

The FDJ sent the same type of email in July 2019 to a random sample of 303,000 online gamblers who had an active gambling account in the monopoly from FDJ. The e-mail contained an invitation to respond to an online survey, with the same content as for the ARJEL survey, hosted by the University Hospital of Nantes. A total of 5,682 gamblers (1.9% of those invited to participate) responded to the whole survey.

Data processing and measures. Both datasets contained basic demographic data (age and sex), gambling tracking data during the twelve months preceding survey completion, and answers to the survey questions (see Table S1 of the supplementary material for a detailed list of variables and how they were operationalized). The list of metrics extracted from the gambling accounts was determined with the objective to have a model that could handle a large roster of gamblers without running into run time or memory issues. As a consequence, data providers eliminated metrics that were found to be computationally infeasible. This was especially the case for time-related metrics; for example, computing session length requires extracting start and end points from a sequence of bets' timestamps and looping the sequence of timestamps multiple times.

Gambling tracking data. Gambling tracking data included information on accounts (number of active accounts and date of creation of each account) and weekly aggregated data representative of the raw gambling activity (e.g., number of bets, deposits, use of loyalty bonuses). Additionally, we computed several types of indicators with potential associations with problematic gambling behavior.

The first type was related to chasing behavior, which is defined as the continuation and/or intensification of gambling after a sequence of losses (Breen & Zuckerman, 1999). The chasing behavior is indeed considered as a very relevant indicator for identifying gamblers at risk of gambling problems, especially based on behavioral data (Breen & Zuckerman, 1999; Challet-Bouju et al., 2020; Ciccarelli, Cosenza, D'Olimpio, Griffiths, & Nigro, 2019; Deng et al., 2019; Toce-Gerstein, Gerstein, & Volberg, 2003). Because the chasing behavior is not observable directly from the gambling tracking data, we had to approximate it with two proxies as was done by Perrot et al. (2017). The first was based on the observation of recurrent deposits of money within a short period of time. The second was defined as making a deposit quickly after placing a bet. A repeated sequence of deposits or a deposit that quickly follows a bet may reflect the case when the gambler lost all the remaining money on his account, and then tries to recover his losses by depositing money back on his account quickly afterwards, beyond his forecasts. Two indicators of chasing were thus computed on a bet-by-bet basis by the ARJEL and the FDJ prior to weekly aggregation, and were operationalized as "making three deposits in less than 12 h" or "making a deposit less than 1 h after placing a bet".

Moreover, we computed a second type of indicator related to breadth of involvement, defined as the range of participation in various forms of gambling. Breadth of involvement is traditionally measured as the number of games an individual plays (Binde, Romild, & Volberg, 2017). High breadth of involvement, also referred to as versatility (Welte, Barnes, Tidwell, & Hoffman, 2009), means that the gambler is engaged in multiple forms of gambling (Binde et al., 2017). It has been found to be associated with problem gambling, potentially as a moderator between gambling on the internet and developing gambling problems (Baggio et al., 2017). We computed the breadth of involvement as the number of different games for which at least one bet was placed by a given participant. This variable ranged from 1 to 10 for the ARJEL and from 1 to 3 for the FDJ (see the online appendix for the types of game considered).

The last type of indicator was related to the longitudinal variability of gambling behavior over time, particularly within individual deviation from "usual" gambling activity (determined from the previous three months). These indicators were calculated *a posteriori* as described by Perrot et al. (2018).

Online survey. The online survey contained questions related to participation in online and offline gambling. We also included the nine items of the PGSI, using a twelve-month time frame. For each positive response to a PGSI item (i.e., a response other than "never"), we duplicated the item, shrinking the time frame from twelve months to thirty days. We calculated twelve-month and thirty-day (i.e., current) PGSI scores, and then used the current scores to classify the participants as non-problem gamblers (0), low-risk gamblers (1–4), moderate-risk gamblers (5–7), and problem gamblers (8 plus) (Currie et al., 2013).

Development of the prediction model

Filtering and analyzing the analytic sample. As the reference period of current PGSI status was the last thirty days, we used data from the previous four months to develop the model and predict current PGSI status. Indeed, all gambling indicators were aggregated at the month level (i.e. over five-week periods) and variability indicators were estimated in relation to the previous three months (usual activity). The four-month period was a trade-off between having a sufficient hindsight of gambling activity, and having a reactive screening tool that would not require going too far back in gambling activity history. As a consequence, we excluded from the analyses gamblers who had created their account less than four months before survey completion (n = 1,134 for the ARJEL dataset and n = 278 for the FDJ dataset). We also excluded individuals who did not gamble in the reference period covered by the current PGSI status (i.e., thirty days before the survey) (n = 813 for the ARJEL dataset and n = 325 for the FDJ dataset). The final ARJEL and FDJ datasets contained 7,359 and 5,079 gamblers, respectively.

Strategy for model testing and selection. We split each dataset into a training sample (80%) and a test sample (20%) and used a two-step strategy.

In the first step, we developed three independent binary models that aim to classify each gambler as problematic or non-problematic based on each of the three thresholds from the PGSI: 1 for low-risk gambling, 5 for moderate-risk gambling and 8 for problem gambling. For each threshold, we tested four machine learning algorithms: random forests, support vector machines, artificial neural networks, and logistic regression. We first performed a repeated (6 times) stratified 5-fold cross-validation with the training sample to build the four algorithms. The parameterization of each machine learning algorithm is described in Table S2 of the supplementary material. We then assessed the predictive performance of each algorithm using several indicators: the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUCPR) (Sofaer, Hoeting, & Jarnevich, 2019), the F1 score (Powers, 2011) at the 0.5 cutoff and at the cutoff maximizing the F1 score, sensitivity when specificity is fixed at 80%, specificity when sensitivity is fixed at 80%, the Youden's index (Fluss, Faraggi, & Reiser, 2005), sensitivity at the cutoff of the Youden's index and specificity at the cutoff of the Youden's index. Finally, we selected the algorithm with the best predictive performance for each binary classification model, based primarily on the maximization of the sensitivity and specificity (i.e., Youden's index). In the case where two or more algorithms resulted in the same Youden's index, we used the F1 score. We used Youden's index to identify the cutoff maximizing the sensitivity and the specificity to classify the gamblers as problematic or not in each binary model.

In the second step, we established a four-class classification based on a nesting of the 3 binary models, i.e. when a gambler was classified as problematic in the PGSI8 binary model, he/she was classified as a problem gambler in the four-class classification, when a gambler was classified as non-problematic in the PGSI8 binary model and problematic in the PGSI5 binary model, he/she was classified as a moderate-risk gambler in the four-class classification, and so on. All the rules are explained in Fig. 1.

The four-class classification was our ultimate goal in order to provide a tool that can be used in real life. Indeed, it may deliver the positioning of each gambler within the four levels of gambling severity directly, thus allowing to provide a tailored feedback to the gambler. However, in contrast to a direct one-step estimation of the four-class model, this two-step approach allows providing three binary classification models that differ regarding the threshold chosen for gambling problems and that can be used either independently or in combination (multiclass). This binary approach with multiple thresholds may be useful to adapt the model used in relation to expectancies from different stakeholders. For example, in a harm-reduction view, public health services or gambling regulation authorities might be more interested in detecting gambling problems from the low-risk level, in order to prevent gambling-related harm, rather than detecting an established gambling disorder, which may be more interesting for researchers or clinicians in a relapse prevention perspective (Browne & Rockloff, 2017; Currie et al., 2009; Delfabbro & King, 2019; Dowling et al., 2021).

Figure 1 summarizes the whole classification process.

For the ARJEL dataset, we included 22 variables in the prediction models: 14 gambling indicators and 8 intra-gambler variability indicators. For the FDJ, we included 15 variables: 14 gambling indicators and the variability indicator related to the number of deposits. The other variability indicators could not be computed because the underlying mixed models did not converge.

We used the test samples to assess the models' performances, based on the AUROC, sensitivities, and specificities. For each of the final four-class models (for the ARJEL and FDJ datasets), we calculated the occurrence of lifetime temporary self-exclusion in each predicted class. Self-exclusion is often used in studies where no validated screening tool is available (Deng et al., 2019).

Moreover, although the prediction models do not aim at identifying potential risk factors, we performed an analysis of the relative importance of gambling indicators in the classification process. We used the randomForestExplainer R package (Paluszynska, Biecek, Jiang, & Jiang, 2017) to

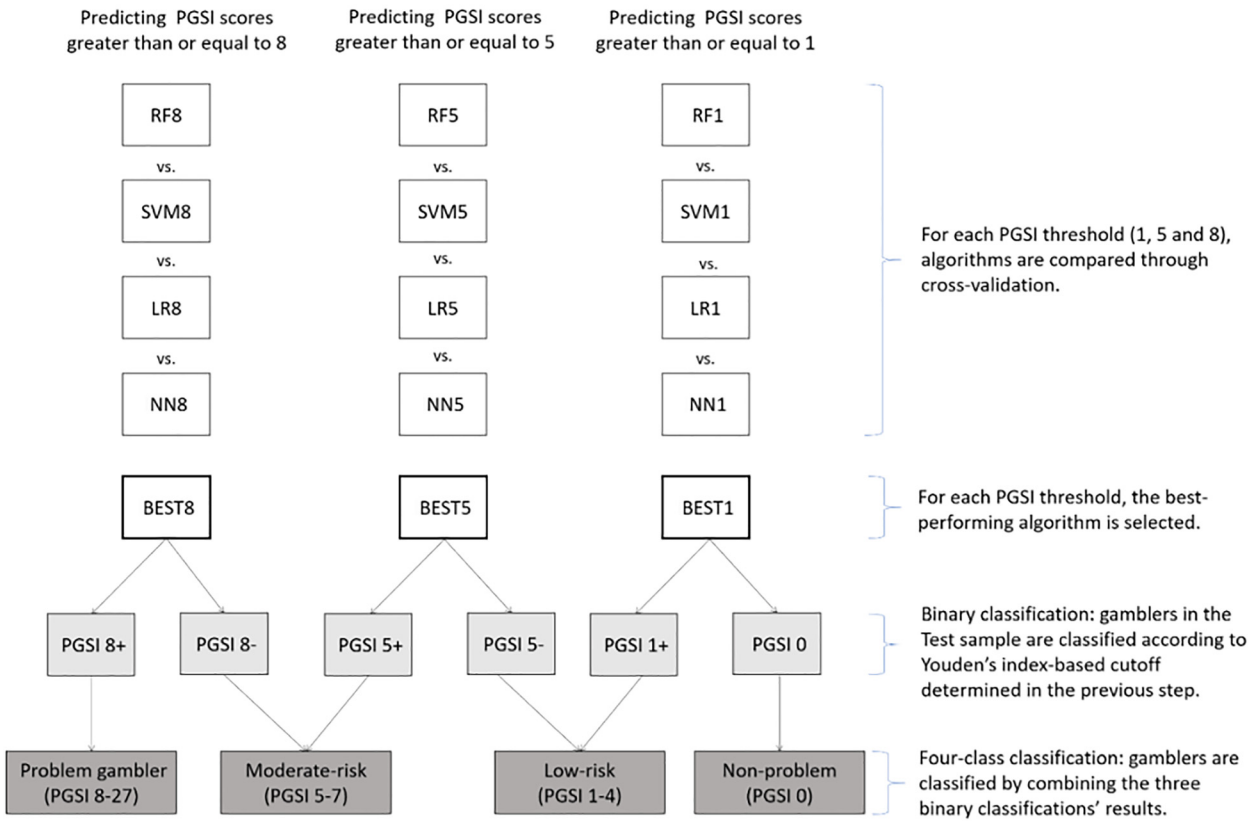

Fig. 1. Summary of the classification process: comparison of algorithms' performance, selection of the best performing algorithm for each PGSI threshold (1, 5 and 8), two-class classification and four-class classification
Notes: RF: random forests; SVM: support vector machines; LR: logistic regression; NN: neural networks

compute six importance measures for each variable: mean decrease in accuracy (how much accuracy the model losses by excluding the considered variable), mean decrease in Gini coefficient (how much the considered variable contributes to the homogeneity of the nodes and leaves in the resulting random forest), number of trees (total number of trees in which the considered variable is used at least once for splitting), mean minimal depth (depth of the node that splits on the considered variable and is the closest to the root of the tree), number of nodes (number of times the considered variable is used for splitting), and times a root (number of trees in which the considered variable is used for splitting the root node). For each importance measure, 1 point was given if it was ranked first, 2 points if it was ranked second, etc. Thus, variables with the lowest total ranking were considered the most important in the classification process.

Finally, we analyzed three possible reasons for misclassification. Reason 1 was based on an item from the online survey: "*In your opinion, which type of gambling has contributed the most to the gambling-related problems you mentioned above?*" (multiple possible answers: Scratch games and lotteries/Sports betting/Horse race betting/Poker/Online casinos). Only gamblers with a PGSI score greater than or equal to 3 answered this question. It allowed us to identify gamblers who did not cite any of the forms of online gambling included in the corresponding dataset (i.e., sports betting, horse race betting, and poker for the ARJEL dataset, and lotteries for the FDJ dataset), who were unlikely to be detected as having gambling problems based on this dataset (potential false negatives). Reason 2 was also based on an item from the online survey: "*Do you gamble…*" (mutually exclusive answers: mostly online/mostly offline/as much online as offline). We used this question to identify false-negative cases that could be explained by participants who gambled primarily offline. Indeed, as data related to land-based gambling was not included in both datasets, these participants were unlikely to be identified as having gambling problems based on these datasets. Finally, Reason 3 was based on the comparison of the twelve-month and the thirty-day PGSI classes. We indeed hypothesized that some false-positive cases could be explained by divergences between the gambling behavior in the four months preceding the survey completion, which are used by the models to estimate the predicted PGSI class, and the timeframe of the current PGSI (past thirty days), especially in case of current gambling abstinence after previously engaging in excessive gambling behavior (potential false positives).

Ethics

This study was approved by the local research ethics committee Groupe Nantais d'Ethique dans le Domaine de la Santé (GNEDS) on March 25, 2015. Participants were informed about the research and gave their written informed consent online prior to their inclusion in the study.

RESULTS

Descriptive statistics

Characteristics and gambling activity over the five weeks preceding survey completion are presented in Table 1 for the ARJEL dataset and in Table 2 for the FDJ dataset. The distribution of PGSI scores (Figures S1 (ARJEL) and S2 (FDJ)) and PGSI items (Tables S5 (ARJEL) and S6 (FDJ)) are provided in the supplementary material. The only variables included in the models but not shown explicitly in these tables are the specific indicators of intraindividual longitudinal variability because they were standardized (i.e., mean = 0 and *sd* = 1) by definition. Certain variables had a very right-skewed distribution, especially those related to amounts of money (e.g. money wagered, withdrawals). This means that the majority of gamblers have low values whereas a small number have higher values, which is common with gambling-tracking data.

Development and validation of the prediction model for the ARJEL dataset

The training sample contained 5,887 gamblers. The test sample contained the remaining 1,472 gamblers. Table 3 shows the predictive performance of the four machine learning algorithms for each PGSI threshold using cross-validation. Based on the estimated performance metrics, we selected random forests as the best performing algorithm for the three PGSI thresholds.

In the test sample, the AUROC for the detection of problem gamblers was 0.82 (CI: [0.78,0.86]), 0.80 (CI: [0.76,0.83]) for moderate-risk gamblers and 0.72 (CI: [0.70,0.75]) for low-risk gamblers. Table 4 provides the 2 × 2 confusion matrices and the sensitivity and specificity for each PGSI threshold and Table 5 shows the results for the four-class classification.

According to Table 5, the final model for the ARJEL dataset correctly identified 71% of non-problem gamblers,

Table 1. Characteristics of participants and description of gambling activity in the five weeks preceding the survey (ARJEL dataset, n = 7,359)

	Categorical variables	Numerical variables						
	n (%)	Mean	SD	Min	P25	Median	P75	Max
Age (years)		44	15	19	32	42	56	96
% of males[a]	6,642 (90%)							
Number of active accounts*		59	51	4	25	52	66	316
Age of the oldest account (months)		1.9	1.5	1.0	1.0	1.0	2.0	24.0
Money wagered (€)**		1,342	7,705	0	39	177	708	279,978
Number of bets**		217	746	1	15	54	166	18,930
Largest single-day total money wagered (€)**		205	959	0	10	33	122	43,162
Losses (€)[b]**		245	2,780	−21,871	3	37	167	203,000
Deposits (€)**		226	793	0	0	35	155	27,300
Number of deposits**		5	10	0	0	1	5	154
Largest single-day total deposits (€)**		65	187	0	0	20	51	4,300
Withdrawals (€)*		132	1,371	0	0	0	0	103,869
Gambling days**		12	11	0	2	9	20	35
Number of different games*		2.5	2.0	1.0	1.0	2.0	3.0	10.0
Loyalty bonuses used (€)*		17	147	0	0	0	3	7,175
Chasing proxy 1[c]*		0.6	3.5	0.0	0.0	0.0	0.0	110.0
Chasing proxy 2[d]*		1.4	4.9	0.0	0.0	0.0	1.0	88.0
Lifetime temporary self-exclusion	630 (8.6%)							
30-days PGSI score		2.0	3.7	0	0	1	2	27
PGSI ≥ 8	544 (7.4%)							
PGSI ≥ 5	1,020 (13.9%)							
PGSI ≥ 1	3,749 (50.9%)							
Non-problem (PGSI 0)	3,610 (49.1%)							
Low-risk (PGSI 1-4)	2,729 (37.1%)							
Moderate-risk (PGSI 5-7)	476 (6.5%)							
Problem gambler (PGSI 8-27)	544 (7.4%)							

[a] Information is missing for n = 81 participants because of conflicting information for multiple-account users.
[b] Positive values indicate losses; negative values indicate wins.
[c] Number of occurrences of the event "Making three deposits in less than 12 h".
[d] Number of occurrences of the event "Making a deposit less than 1 h after placing a bet".
* Variable included in the prediction models.
** Variable included in the prediction models and for which the corresponding variability indicator was also included in the prediction models.
P25: 25th percentile; P75: 75th percentile.

Table 2. Characteristics of participants and description of gambling activity in the five weeks preceding the survey (FDJ dataset, $n = 5{,}079$)

	Categorical variables *n* (%)	Numerical variables						
		Mean	SD	Min	P25	Median	P75	Max
Age (years)		53	13	21	42	52	62	80
% of males[a]	3,292 (65%)							
Age of account		85	54	13	34	76	131	221
Money wagered (€)*		92	280	0.25	22	42	82	7,131
Number of "bets"*		34	171	1	3	9	19	5,614
Losses (€)[a]*		33	804	−49,639	14	30	58	2,742
Largest single-day total money wagered (€)*		24	48	0	6	12	26	1,274
Deposits (€)*		54	123	0	10	25	55	2,803
Number of deposits**		2	7	0	1	1	3	309
Largest single-day total deposits (€)*		25	31	0	10	25	25	500
Withdrawals (€)*		22	808	0	0	0	0	49,954
Gambling days*		7	7	1	2	5	10	35
Loyalty bonuses used (€)*		0.2	0.9	0.0	0.0	0.0	0.0	13.5
Number of different games*		1.3	0.6	1	1	1	2	3
Number of gambling moderator modifications*		0.03	0.24	0.00	0.00	0.00	0.00	4.00
Chasing proxy 1[b]*		0	1	0	0	0	0	33
Chasing proxy 2[c]*		0	5	0	0	0	0	253
Lifetime temporary self-exclusion	49 (1%)							
30-day PGSI score		0.5	1.4	0.0	0.0	0.0	0.0	19.0
PGSI ≥8	41 (0.8%)							
PGSI ≥5	115 (2.3%)							
PGSI ≥1	981 (19.3%)							
Non-problem (PGSI 0)	4,098 (80.7%)							
Low-risk (PGSI 1–4)	866 (17.1%)							
Moderate-risk (PGSI 5–7)	74 (1.5%)							
Problem gambler (PGSI 8–27)	41 (0.8%)							

[a] Positive values indicate losses; negative values indicate wins.
[b] Number of occurrences of the event "Making three deposits in less than 12 h".
[c] Number of occurrences of the event "Making a deposit less than 1 h after placing a bet".
* Variable included in the prediction models.
** Variable included in the prediction models and for which the corresponding variability indicator was also included in the prediction models.
P25: 25th percentile; P75: 75th percentile.

18% of low-risk gamblers, 7% of moderate-risk gamblers, and 75% of problem gamblers.

Among the gamblers classified in the non-problem category, 4% (26/709) had set up at least one temporary self-exclusion since they started gambling online, compared to 8% (45/546) in the low-risk category, 31% (33/107) in the moderate-risk category, and 38% (42/110) in the problem gambling category.

After exploration of the possible causes for misclassification, 24 participants were excluded from the confusion matrix due to Reason 1 (type of gambling), 182 due to Reason 2 (offline gambling), and 102 due to Reason 3 (twelve-month vs thirty-day PGSI status). The corresponding "corrected" confusion matrix is provided in Table S3 from the supplementary material, and shows an improvement in the proportion of gamblers correctly identified: 84% (505/599) for non-problem gamblers, 24% (98/401) for low-risk gamblers, 12% (8/66) for moderate-risk gamblers, and 85% (83/98) for problem gamblers.

Development and validation of the prediction model for the FDJ dataset

The training sample contained 4,423 gamblers. The test sample contained the remaining 1,016 gamblers. Table 6 shows the predictive performance of the four machine learning algorithms for each PGSI threshold. We selected random forest as the best performing algorithm for the three PGSI thresholds.

In the test sample, the AUROC for the detection of problem gamblers was 0.76 (CI: [0.59,0.93]), 0.75 (CI: [0.64,0.85]) for moderate-risk gamblers and 0.63 (CI: [0.59,0.68]) for low-risk gamblers. Table 7 provides the 2×2 confusion matrices and the sensitivity and specificity for each PGSI threshold, and Table 8 shows the results for the four-class classification.

According to Table 8, the final model for the FDJ dataset correctly identified 68% (566/835) of non-problem gamblers, 23% (37/158) of low-risk gamblers, 0% (0/12) of moderate-risk gamblers, and 55% (6/11) of problem gamblers.

Table 3. Cross-validated predictive performance measures of the binary classification algorithms in the training sample of the ARJEL dataset (n = 5,887)

	Predicting PGSI ≥8			
	RF	SVM	LR	NN
AUROC	**0.83**	0.59	0.80	0.71
AUCPR	**0.33**	0.20	0.29	0.20
F1	**0.13**	-	0.12	-
F1 max	**0.41**	0.32	0.39	0.29
Specificity when sensitivity = 80%	**0.71**	0.18	0.69	0.25
Sensitivity when specificity = 80%	**0.71**	0.40	0.70	0.29
Youden's index	**0.54**	0.29	0.52	0.40
Sensitivity at Youden's cutoff	**0.76**	0.40	0.75	0.74
Specificity at Youden's cutoff	0.78	**0.89**	0.77	0.66
	Predicting PGSI ≥ 5			
AUROC	**0.80**	0.68	0.78	0.78
AUCPR	**0.45**	0.38	0.40	0.40
F1	**0.35**	0.17	0.25	-
F1 max	**0.48**	0.46	0.47	0.47
Specificity when sensitivity = 80%	**0.64**	0.29	0.62	0.60
Sensitivity when specificity = 80%	**0.66**	0.58	0.64	0.63
Youden's index	**0.49**	0.41	0.47	0.48
Sensitivity at Youden's cutoff	**0.73**	0.57	0.71	**0.73**
Specificity at Youden's cutoff	0.75	**0.85**	0.76	0.74
	Predicting PGSI ≥ 1			
AUROC	**0.70**	0.69	0.69	**0.70**
AUCPR	**0.71**	0.68	**0.71**	**0.71**
F1	**0.64**	0.62	0.60	0.63
F1 max	**0.69**	**0.69**	0.68	**0.69**
Specificity when sensitivity = 80%	0.43	**0.45**	0.41	0.44
Sensitivity when specificity = 80%	**0.50**	0.48	**0.50**	0.48
Youden's index	**0.32**	0.31	0.31	**0.32**
Sensitivity at Youden's cutoff	0.56	**0.62**	0.56	0.61
Specificity at Youden's cutoff	**0.76**	0.70	**0.76**	0.71

Notes: RF: random forests; SVM: support vector machines; LR: logistic regression; NN: neural networks; AUROC: area under the receiver operating characteristic curve; AUCPR: area under the precision-recall curve; F1: F1 score; F1 max: best F1 score that could be obtained by selecting the optimal cutoff.
Bold values indicate the best predictive performance amongst the four algorithms.

Among gamblers classified in the non-problem category, 0.5% (4/835) had set up at least one temporary self-exclusion, compared to 1% (2/158) in the low-risk category, 8% (1/12) in the moderate-risk category and 9% (1/11) in the problem gambling category.

After accounting for potential reasons for misclassification, we excluded 49 gamblers due to Reason 1 (type of gambling), 53 due to Reason 2 (offline gambling), and 32 due to Reason 3 (twelve-month vs thirty-day PGSI status). The corresponding confusion matrix is provided in Table S4 from the supplementary material, and shows an improvement in the proportion of gamblers correctly identified: 73% (566/771) for non-problem gamblers, 27% (37/136) for low-risk gamblers, 0% (0/6) for moderate-risk gamblers, and 67% (6/9) for problem gamblers.

Assessment of variables' importance

For both datasets, we assessed the relative "importance" of each gambling indicator and for each PGSI threshold to obtain insight into how the random forests algorithm classifies the individuals as non-problem or problem gamblers. Figures 2 and 3 show that the most important gambling indicators differ according to gambling type (pure chance games vs. skill-based games). For the ARJEL dataset, the importance of the variables also varies according to the PGSI thresholds, especially for deposit-related indicators.

DISCUSSION

Gamblers included in the two datasets were quite similar to the source population of French online gamblers (J.-M. Costes & Eroukmanoff, 2018), i.e. a higher proportion of males for sports betting, horse race betting and poker (72–82% in the survey, 90% in our sample) compared to lotteries (61% in the survey, 65% in our sample), and a higher age for lotteries gamblers and horse race bettors compared to sports bettors and poker players. These gambling type-related demographic characteristics have also been found in other online gambler populations from other countries (McCormack, Shorter, & Griffiths, 2014; Wood & Williams, 2011, p. 27). Moreover, observed frequencies of problem gamblers in both datasets (7.4% for the ARJEL dataset and 0.8% for the FDJ dataset) seem to be very close to those observed in the last French prevalence survey (5.9%, 6.2%, 4.0% and 0.8% for sports betting, horse race betting, poker and lotteries, respectively) (J.-M. Costes et al., 2020), so the samples used seem to be representative of the source population. The prevalence of problem gamblers was also comparable to the results found in other surveys among online gamblers (Macey & Hamari, 2018; Price, 2022; Tomei, Petrovic, & Simon, 2022).

Summary of predictive performance for the ARJEL dataset

The AUROCs of the three binary classification models (i.e., when each PGSI threshold is considered separately) ranged from 0.72 to 0.82 according to the PGSI threshold.

Table 4. Confusion matrix for each PGSI threshold in the test sample of the ARJEL dataset (n = 1,472)

	Observed PGSI <8	Observed PGSI ≥8	Observed PGSI <5	Observed PGSI ≥5	Observed PGSI <1	Observed PGSI ≥1
Classified as negative	1,062	27	939	69	531	321
Classified as positive	300	83	316	148	178	442
Specificity/sensitivity	78%	76%	75%	68%	75%	58%

Table 5. Confusion matrix of the four-class classification in the test sample of the ARJEL dataset (n = 1,472)

	Observed PGSI class				
Predicted PGSI class	Non-problem (PGSI = 0)	Low-risk (PGSI 1–4)	Moderate-risk (PGSI 5–7)	Problem gambler (PGSI 8–27)	Total
Non-problem (PGSI 0)	**505**	248	29	14	796
Low-risk (PGSI 1–4)	75	**98**	14	8	195
Moderate-risk (PGSI 5–7)	37	48	**8**	5	98
Problem gambler (PGSI 8–27)	92	152	56	**83**	383
Total	709	546	107	110	1,472

Regarding the final four-class classification model derived from the three binary models, we can assess its performance by analyzing the confusion matrix in the test sample and considering the clinical meaning of the categories and the proximity between adjacent categories.

Table 6. Cross-validated predictive performance measures of the binary classification algorithms in the training sample of the FDJ dataset (n = 4,423)

	Predicting PGSI ≥8			
	RF	SVM	LR	NN
AUROC	**0.74**	0.50	0.68	0.52
AUCPR	0.15	**0.16**	0.15	0.10
F1	-	-	-	-
F1 max	0.16	**0.18**	0.17	0.02
Specificity when sensitivity = 80%	0.31	0.27	**0.50**	0.06
Sensitivity when specificity = 80%	**0.62**	0.34	0.52	0.05
Youden's index	**0.53**	0.29	0.47	0.07
Sensitivity at Youden's cutoff	0.67	0.58	0.70	**0.89**
Specificity at Youden's cutoff	**0.86**	0.71	0.77	0.18
	Predicting PGSI ≥ 5			
AUROC	**0.77**	0.51	0.71	0.54
AUCPR	0.12	0.10	**0.15**	0.07
F1	-	-	-	-
F1 max	0.21	0.17	**0.23**	0.08
Specificity when sensitivity = 80%	**0.51**	0.15	0.42	0.08
Sensitivity when specificity = 80%	**0.60**	0.32	0.54	0.09
Youden's index	**0.47**	0.22	0.42	0.08
Sensitivity at Youden's cutoff	**0.70**	0.37	0.66	0.18
Specificity at Youden's cutoff	0.78	0.86	0.77	**0.90**
	Predicting PGSI ≥ 1			
AUROC	**0.65**	0.60	**0.65**	0.64
AUCPR	**0.36**	0.35	**0.36**	0.35
F1	**0.21**	0.09	0.14	-
F1 max	**0.40**	0.39	**0.40**	0.39
Specificity when sensitivity = 80%	**0.37**	0.21	0.33	0.33
Sensitivity when specificity = 80%	0.40	**0.42**	**0.42**	0.37
Youden's index	**0.25**	0.24	**0.25**	0.23
Sensitivity at Youden's cutoff	**0.55**	0.39	**0.55**	0.51
Specificity at Youden's cutoff	0.70	**0.85**	0.70	0.72

Notes: RF: random forests; SVM: support vector machines; LR: logistic regression; NN: neural networks; AUROC: area under the receiver operating characteristic curve; AUCPR: area under the precision-recall curve; F1: F1 score; F1 max: best F1 score that could be obtained by selecting the optimal cutoff.
Bold values indicate the best predictive performance amongst the four algorithms.

Among problem gamblers, the majority were identified by the model as such. Regarding moderate-risk gamblers, only a small proportion were correctly identified. Among the misclassified gamblers, two-thirds were classified as problem gamblers. This result suggests difficulty in distinguishing between moderate-risk and problem gamblers. However, from a responsible gambling perspective, it may be preferable to classify a moderate-risk gambler as a problem gambler rather than as a low-risk or non-problem gambler. For the low-risk gamblers, almost one quarter were correctly classified. Close to half of misclassified gamblers (43%) were classified as non-problem gamblers, which can be considered acceptable based on the proximity between the two categories and the low level of risk related to these two categories. What is more concerning is the proportion classified as problem (24%) or moderate-risk (8%) gamblers. These misclassified gamblers could thus be considered "full" false positives. Regarding the non-problem gamblers, the large majority were correctly identified. Misclassified gamblers were mainly classified in the upper class (low-risk gamblers, 6%), which, once again, is preferable from a responsible gambling perspective. Other misclassified gamblers were classified as moderate-risk gamblers (3%) and problem-gamblers (6%) and could be considered "full" false positives.

Summary of the predictive performance for the FDJ dataset

Interpreting the performance results for the FDJ dataset requires caution because of the low prevalence of problem gambling in the sample (2.3% of moderate-risk gamblers and 0.8% of problem gamblers). Indeed, this low prevalence, which was expected, results in wide confidence intervals around the AUROCs, especially for the PGSI 8 threshold (0.76 [95% CI: 0.59–0.93]) and the PGSI 5 threshold (0.75 [95% CI: 0.64–0.85]). As for the ARJEL dataset, the four-class confusion matrix shows that non-problem and problem gamblers were mainly identified as such (55%–68%), with misclassified gamblers mainly identified in the adjacent categories. However, only a small proportion (0%–23%) of low-risk and moderate-risk gamblers were correctly identified. Among them, more than a third of misclassified low-risk gamblers were "full" false positives (classified as moderate-risk and problem

Table 7. Confusion matrix for each PGSI threshold in the test sample of the FDJ dataset (n = 1,016)

	Observed PGSI <8	Observed PGSI ≥8	Observed PGSI <5	Observed PGSI ≥5	Observed PGSI <1	Observed PGSI ≥1
Classified as negative	915	5	849	12	598	94
Classified as positive	90	6	144	11	237	87
Specificity/Sensitivity	91%	55%	86%	48%	72%	48%

Table 8. Confusion matrix of the four-class classification in the test sample of the FDJ dataset (n = 1,016)

	Observed PGSI class				
Predicted PGSI class	Non-problem (PGSI = 0)	Low-risk (PGSI 1–4)	Moderate-risk (PGSI 5–7)	Problem gambler (PGSI 8–27)	Total
Non-problem (PGSI 0)	**566**	79	7	1	653
Low-risk (PGSI 1–4)	138	**37**	2	2	179
Moderate-risk (PGSI 5–7)	70	16	**0**	2	88
Problem gambler (PGSI 8–27)	61	26	3	**6**	96
Total	835	158	12	11	1,016

gamblers) and more than half of misclassified moderate-risk gamblers were "full" false negatives (classified as non-problem gamblers). Changing the cutoff probabilities could increase the sensitivity of the model at the cost of decreasing its specificity, and *vice versa*. However, obtaining a good trade-off between sensitivity and specificity would remain difficult because of the low prevalence of problem gamblers in the FDJ sample.

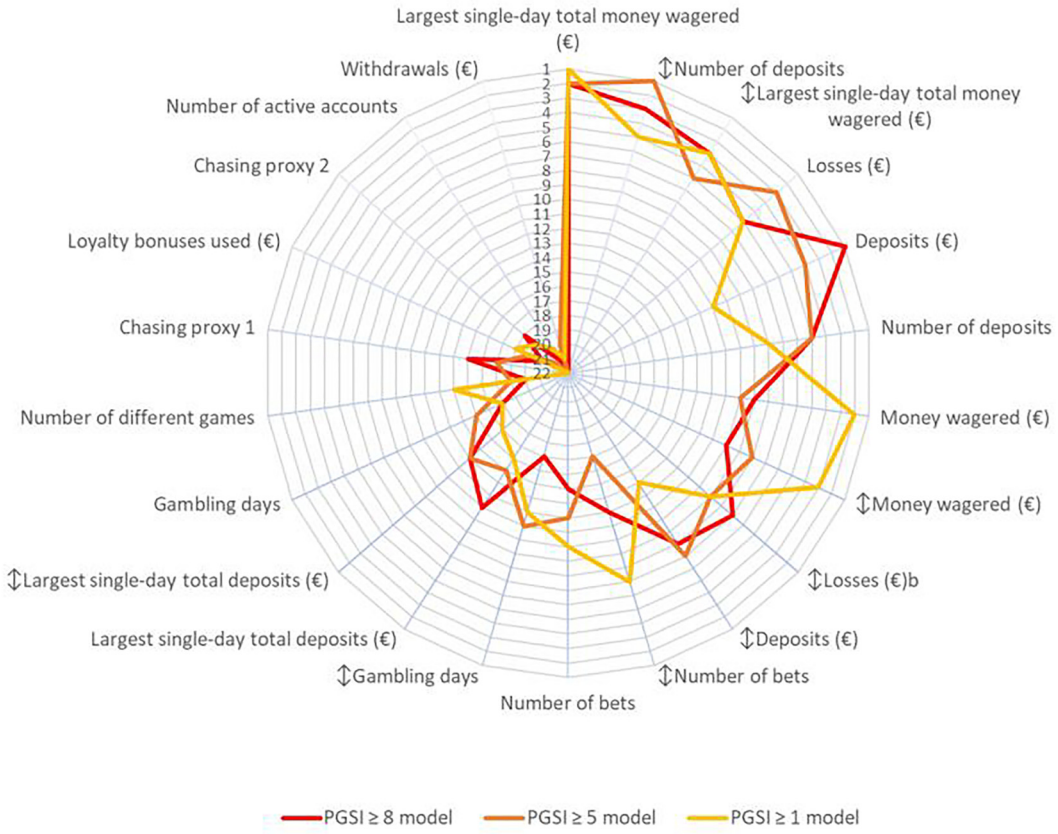

Fig. 2. Most important variables for each PGSI threshold according to variables' ranking (from 1: most important to 22: less important) computed from six importance measures (ARJEL dataset)
Notes: ↕ denotes indicators of longitudinal variability

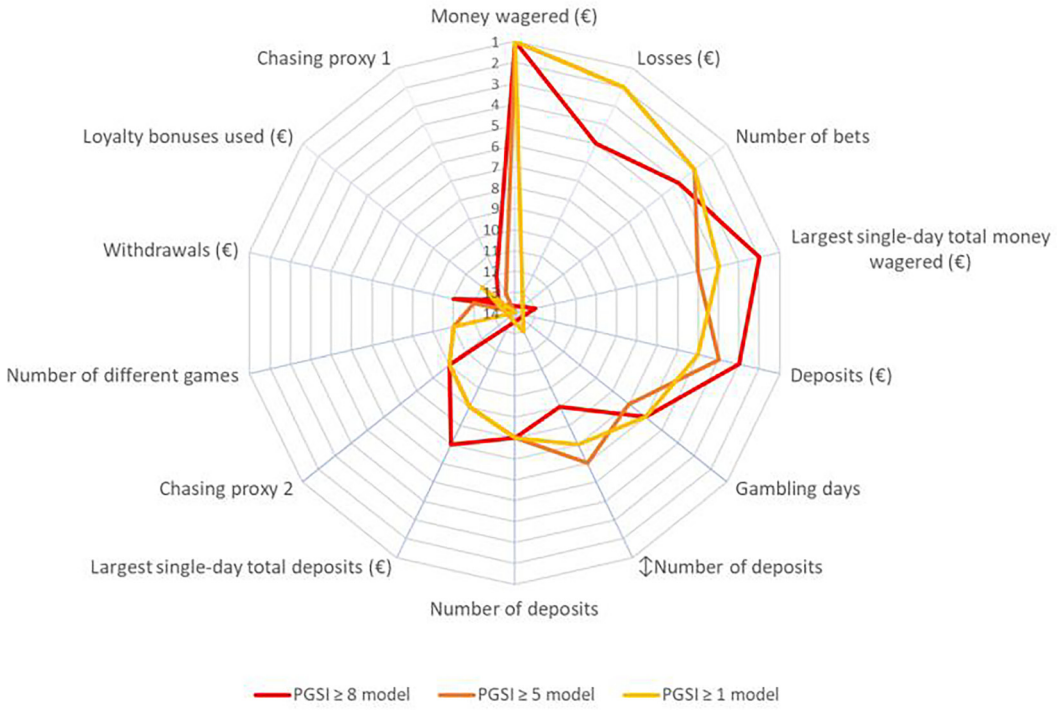

Fig. 3. Most important variables for each PGSI threshold according to variables' ranking (from 1: most important to 15: less important) computed from six importance measures (FDJ dataset)
Notes: ↕ denotes indicators of longitudinal variability

Relative importance of gambling indicators

Regarding "variables' importance" in the models fit using the ARJEL data, it appears that deposit-related variables are important for predicting problem and moderate-risk gamblers. Regarding the detection of low-risk gamblers, money wagered (total amount and large amounts wagered in a single day) and number of bets were more important variables than deposit-related indicators. Thus, the amount of money wagered can be useful to discriminate between non-problem and other gamblers but is not a sufficient indicator to detect individuals with the most severe levels of gambling problems (Nelson et al., 2008); deposit-related variables appear more useful in that case. Indeed, deposits reflect the actual expenses of the gamblers (i.e. money that comes out of their bank accounts) whereas money wagered may be more the reflect of the level of participation in gambling. At a certain level of gambling problems (i.e. at least moderate-risk gambling), the majority of gamblers may have a high participation in gambling but those with more severe gambling problems may experience more expenses (i.e. more losses). Breadth of involvement (measured as the number of different games played) does not seem to be an important indicator according to its ranking, although it is considered a distinguishing characteristic of problem gamblers (Deng et al., 2019; S. M. Gainsbury, 2015). This might be because in the present study, the number of different games was computed without information on non-regulated online gambling activity (e.g. online casinos) and land-based gambling. Moreover, the level of detail was lower compared to other studies that focused on breadth of involvement and considered a finer granularity of gambling types (LaPlante, Nelson, & Gray, 2014). Additionally, the two proxies of chasing behavior do not appear to be very important variables, although they have known potential to discriminate between social and problem gamblers (Ciccarelli et al., 2019; Deng et al., 2019; Temcheff, Paskus, Potenza, & Derevensky, 2016). We can hypothesize that the two proxies we used were more "specific" indicators than "sensible" indicators and thus appeared moderately important in a sample with a low percentage of problem gamblers. Had we performed direct one-on-one comparisons, especially by trying to discriminate between moderate-risk and problem gamblers, these indicators would probably have been more central. Moreover, several structural characteristics of online games may facilitate chasing behavior, such as in-running betting, payout ratio (McCormack & Griffiths, 2013), or interval ratio reinforcement schedule of conditioning (Lister, Nower, & Wohl, 2016). Our study focused only on the range of gambling types that are legal online in France (i.e. poker, horse-race betting, sports betting and lotteries). Thus, it excluded online casino games, especially electronic gaming machines, which are typically characterized by very high event frequency (McCormack & Griffiths, 2013). This may also explain why the chasing proxies were not as important as other indicators.

For the pure chance games of the FDJ dataset, money wagered was the most important gambling indicator for all three PGSI thresholds. There was no marked difference in variable importance according to the PGSI threshold.

Implications of the analysis of misclassified gamblers

The confusion matrices obtained after excluding misclassified gamblers according to self-declared problematic gambling forms, offline gambling and twelve-month PGSI provide an estimate of the lack of sensitivity and specificity that could be due to data limitations rather than model-building limitations. Indeed, these "corrected" results correspond to what might be observed if the models were applied to a dataset that included the complete online gambling activity and/or offline gambling data. In 2020, the ARJEL was transformed into a broader national authority, the ANJ (National Gambling Authority), which regulates all forms of gambling, including offline gambling and lotteries. A perspective of this work is thus to apply the algorithm developed to screen for problem gamblers using data related to the entirety of legal offline and online gambling activity.

Limitations and strengths

The model development involved comparing the predictive performance of four supervised classification algorithms. Other algorithms could have been tested, and the ones used in this study could have been parameterized in other ways. The values of sensitivity and specificity might seem rather low, but this range of values are quite common when using behavioral gambling tracking data to identify problem gamblers (Luquiens et al., 2016; Philander, 2014). Moreover, the values observed for the lowest PGSI thresholds (1 and 5) are lower than for the highest (8) threshold likely because individuals with less severe gambling problems are more difficult to be detected as problematic gamblers, and *vice versa*. Access to or inclusion of other variables might have improved our models' accuracy. Additionally, we chose not to include age and gender in the models. Indeed, even if age and gender, especially the latter, are likely to be associated with preferences regarding types of games, we chose to base our algorithms entirely on account-based gambling data, which are modifiable behaviors. This choice is linked with the possible use of these algorithms in the future and the possible prevention interventions that could be connected with (e.g. personalized feedback). Another limitation of the model may be the gold-standard used (i.e. the PGSI). Indeed, self-reported scales have been widely criticized in gambling research, due to the lack of accuracy and validity of the responses reported by the gamblers, especially regarding differences between claimed and actual behavior (Baumeister, Vohs, & Funder, 2007; Braverman et al., 2014; S. Gainsbury, 2011; Garber, Nau, Erickson, Aikens, & Lawrence, 2004; Heirene et al., 2021; Rundle-Thiele, 2009). However, the PGSI was used to measure problem gambling, which is by definition necessarily based on participant-reported outcome, rather than gambling behavior. Moreover, the use of the PGSI represents a good trade-off between scientific relevance and resource saving in high-scale epidemiological study for measuring gambling problems. Regarding the samples used, participants included in the study, in both datasets, were those who accepted completing the survey. As a consequence, the samples suffered from selection biases. Finally, our model is based on legal online gambling activity in France and thus does not account for illegal gambling activity, especially online casino games other than poker.

However, in comparison to the findings of Luquiens et al. (2016) that focused on online poker players (also in a French sample), our model displays slightly better predictive performance. The AUROC of our model for detecting moderate-risk gamblers (the same threshold used by Luquiens et al.) was greater (0.80 vs. 0.73). When fixing the sensitivity at 80%, Luquiens et al. obtained a specificity of 50%, compared to 61% in our model. Moreover, our model also displayed better performance for predicting problem gamblers (PGSI threshold of 8) than in a recent study from Auer and Griffiths (2022) that focused on more than 1,200 players from a European online gambling casino (AUROC of 0.82 and 0.76 vs. 0.73). It is likely that the nature of the data used in our study (gambling tracking data, not focusing on a single gambling operator but rather extended to the whole online authorized gambling activity in France) explains the better prediction performances of our model, because they are more representative of the global activity of the players. Moreover, the distinction between skill-based and pure chance games represent important strengths and provide originality to this work compared to previous literature. Finally, the strategy of using three binary classification models based on the three PGSI thresholds to construct the four-class classification allows to use them independently from each other to fit with different stakeholders needs.

Perspectives

There could be several ways of improving the predictive performance of such model in future research. First, regarding the statistical methods used, an interesting solution for improving the model would be the use of ensemble methods, such as boosting or stacking (Zhou, 2012). These methods consist of combining the predictions of several algorithms, theoretically resulting in better predictions. However, the estimation procedure would be more complex, more black-box like, and more time consuming. We can also note that in the study by Auer and Griffiths (2022), the boosting approach performed worse than Random Forests. Second, regarding the indicators used, accessing to finer gambling data (such as timestamped data that allow for computed time-related metrics) or calculating other proxies possibly indicative of a gambling problem (such as for example the "regular gambling account depletion" indicator calculated in the study by Auer and Griffiths (2022)) may have improved the model. Finally, and probably more importantly, expanding the scope of the data used, both gambling-tracking data (e.g. by also including illegal and offline activity) but also other informative data related to the player (such as his level of income for example), may help to have a more representative view of the global activity and situation

of the players, and thus to predict their risk for having gambling problems.

The ongoing third stage of the EDEIN project, which involves clinical interviews of a subset of gamblers, could help improve the accuracy of the algorithm and provide insight into reasons for misclassification, particularly by studying the relationship between the PGSI classification and the clinical diagnosis of gambling disorders. Furthermore, Stage 3 will allow access to merged gambling tracking data from both the ARJEL and the FDJ for each participant and to complementary data from each gambler's account history (e.g., time-related data). This could result in finding complementary gambling indicators to detect problem gamblers or potentially redefining some important indicators differently, especially those related to chasing behavior and breadth of involvement. Indeed, being able to operationalize these two gambling behavior characteristics with players' account data would likely help identify at-risk gamblers, as both have been linked to problem gambling (Baggio et al., 2017; Deng et al., 2019; LaPlante et al., 2011).

The development of a screening model to detect problem gamblers based on behavioral tracking data could be used for two purposes. The first is the identification itself. By revealing to a gambler his or her specific pattern of gambling, which may be misestimated by the gambler (Drosatos et al., 2020), and positioning the gambler in relation to other gamblers of the same age or gender, the gambler can become aware of the excessive nature of his or her gambling activity, when this is the case. This identification step is a key process in behavioral change. Indeed, according to the transtheoretical approach (Prochaska & DiClemente, 2005), switching between the first change stages (precontemplation → contemplation) requires that the person be able to see the problem, to become aware that the problem exists. By increasing information on one's own behavior, confronting the gambler with the reality of his or her behavior and pointing out those behaviors that are excessive, an identification algorithm may help the gambler accomplish one of the first processes of change, namely consciousness raising. The second purpose of a screening algorithm is to provide interventions for gamblers who are identified as having problems or being at risk of having problems. Such interventions could take different forms, including providing personalized feedback (Harris & Griffiths, 2017), counseling for the optimal use of responsible gambling tools, valuing help-seeking and increasing awareness of help available for gambling problems, providing easy access to a range of treatment options, including self-help programs, that can be suited to a large panel of gamblers profiles (H. Miller, 2014), and providing information on gambling-related harm. The objective of those strategies is to empower the gambler to be an actor in changing his or her behavior when necessary (Drosatos et al., 2020). Moreover, as highlighted by Haefeli et al. (2011), prevention measures are more efficient if they are individualized and occur in the early stages of gambling problems. Rapidly identifying individuals who are at risk for future and/or more severe gambling problems would allow to the development of tailored graduated interventions. Regarding operational issues, Drosatos et al. (2018) proposed a conceptual architecture for a responsible gambling information system, including a "predict behavior" component.

Conclusion

In conclusion, we have developed and internally validated two prediction models for screening problem gamblers based on players' account data. The first screening model was built using data from multiple operators and exhibited good predictive performance, especially for detecting problem gamblers. The second, based on data from a single operator providing online lotteries, showed moderate results and would need to be applied to a larger dataset with a greater number of problem gamblers to be assessed more reliably. Globally, these two prediction models could be useful tools to identify priority gamblers for targeted prevention measures. Gambling authorities or operators could, for instance, recommend the use of gambling moderators, self-exclusion, or specialized care programs for the most problematic gamblers.

Funding sources: This research project has received funding from the Primary Prevention Call for Proposals that was issued by the French Institute for Public Health Research (IREsP) and the French National Cancer Institute (INCa) in 2013 and funded by the French National Health Insurance Fund for Employees, the French Directorate General of Health, the Arc Foundation for Cancer Research, the French National Cancer Institute, the French National Institute for Prevention and Education in Health, the French National Institute of Health and Medical Research, the French Inter-Departmental Agency for the Fight against Drugs and Addictive Behaviors, and the French Social Security Scheme for Liberal Professionals. Neither the data providers (ARJEL/ANJ and FDJ) nor the funders (IREsP and its partners) had any influence on this work; they did not review or approve the manuscript. No constraints on publication were imposed.

Authors' contribution: BP: Analysis and interpretation of data; statistical analysis; methodological conceptualization; wrote first draft of the manuscript. JBH: Methodological conceptualization. ET: Data collection. AS: Data collection. MG-B: Obtained funding; study supervision. GC-B: Study concept and design; analysis and interpretation of data; obtained funding; study supervision; wrote first draft of the manuscript. All authors had full access to all data and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors gave feedback on the first draft of the manuscript and approved the final manuscript.

Conflict of interest: BP and JB-H declare that they have no competing interests in relation to this work. ET, AS, MG-B

and GC-B declare that the University Hospital of Nantes received funding from the gambling industry (FDJ and PMU) in the form of a philanthropic sponsorship (donations that do not assign purpose of use). Scientific independence with respect to these gambling industries is guaranteed, and this funding has never had any influence on the present work.

Acknowledgement: We would like to warmly thank the ARJEL (Autorité de Régulation des Jeux En Ligne), which is newly known as the ANJ (Autorité Nationale des Jeux) since the gambling law reform in 2020, and the FDJ (Française des Jeux) for giving us access to gambling tracking data.

This research was conducted at the initiative of and coordinated by the UIC Psychiatrie et Santé Mentale of the University Hospital of Nantes (CHU Nantes), which sponsored this study.

SUPPLEMENTARY MATERIALS

Supplementary data to this article can be found online at https://doi.org/10.1556/2006.2022.00063.

REFERENCES

Auer, M., & Griffiths, M. D. (2016). Should voluntary "self-exclusion" by gamblers be used as a proxy measure for problem gambling? *MOJ Addiction Medicine & Therapy*, *2*(2), 31–33, 00019. https://doi.org/10.15406/mojamt.2016.02.00019.

Auer, M., & Griffiths, M. D. (2022). Using artificial intelligence algorithms to predict self-reported problem gambling with account-based player data in an online casino setting. *Journal of Gambling Studies*. https://doi.org/10.1007/s10899-022-10139-1.

Baggio, S., Dupuis, M., Berchtold, A., Spilka, S., Simon, O., & Studer, J. (2017). Is gambling involvement a confounding variable for the relationship between Internet gambling and gambling problem severity? *Computers in Human Behavior*, *71*, 148–152. https://doi.org/10.1016/j.chb.2017.02.004.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x.

Binde, P., Romild, U., & Volberg, R. A. (2017). Forms of gambling, gambling involvement and problem gambling: Evidence from a Swedish population survey. *International Gambling Studies*, *17*(3), 490–507. https://doi.org/10.1080/14459795.2017.1360928.

Bjerg, O. (2010). Problem gambling in poker: Money, rationality and control in a skill-based social game. *International Gambling Studies*, *10*(3), 239–254. https://doi.org/10.1080/14459795.2010.520330.

Braverman, J., Tom, M. A., & Shaffer, H. J. (2014). Accuracy of self-reported versus actual online gambling wins and losses. *Psychological Assessment*, *26*(3), 865–877. https://doi.org/10.1037/a0036428.

Breen, R. B., & Zuckerman, M. (1999). 'Chasing' in gambling behavior: Personality and cognitive determinants. *Personality and Individual Differences*, *27*(6), 1097–1111. https://doi.org/10.1016/S0191-8869(99)00052-5.

Browne, M., & Rockloff, M. J. (2017). The dangers of conflating gambling-related harm with disordered gambling: Commentary on: Prevention paradox logic and problem gambling (Delfabbro & King, 2017). *Journal of Behavioral Addictions*, *6*(3), 317–320. https://doi.org/10.1556/2006.6.2017.059.

Catania, M., & Griffiths, M. D. (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*. https://doi.org/10.1007/s10899-021-10080-9.

Challet-Bouju, G., Hardouin, J.-B., Thiabaud, E., Saillard, A., Donnio, Y., Grall-Bronnec, M., & Perrot, B. (2020). Modeling early gambling behavior using indicators from online lottery gambling tracking data: Longitudinal analysis. *Journal of Medical Internet Research*, *22*(8), e17675. https://doi.org/10.2196/17675.

Ciccarelli, M., Cosenza, M., D'Olimpio, F., Griffiths, M. D., & Nigro, G. (2019). An experimental investigation of the role of delay discounting and craving in gambling chasing behavior. *Addictive Behaviors*, *93*, 250–256. https://doi.org/10.1016/j.addbeh.2019.02.002.

Costes, J.-M., & Eroukmanoff, V. (2018). Les pratiques de jeux d'argent sur Internet en France en 2017. 8.

Costes, J.-M., Eroukmanoff, V., Richard, J.-B., & Tovar, M.-L. (2015). Les jeux d'argent et de hasard en France en 2014. *Notes de l'ODJ n°6*, 6. http://www.academia.edu/12112626/Les_jeux_dargent_et_de_hasard_en_France_en_2014.

Costes, J. M., Pousset, M., Eroukmanoff, V., Le Nezet, O., Richard, J. B., Guignard, R., … Arwidson, P. (2011). Les niveaux et pratiques des jeux de hasard et d'argent en 2010. *Tendances*, *77*, 1–8.

Costes, J.-M., Richard, J.-B., & Eroukmanoff, V. (2020). Les problemes lies aux jeux d'argent en France, en 2019. *Les notes de l'Observatoire des Jeux*, *12*, 7.

Currie, S. R., Hodgins, D. C., & Casey, D. M. (2013). Validity of the problem gambling severity index interpretive categories. *Journal of Gambling Studies*, *29*(2), 311–327. https://doi.org/10.1007/s10899-012-9300-6.

Currie, S. R., Miller, N., Hodgins, D. C., & Wang, J. (2009). Defining a threshold of harm from gambling for population health surveillance research. *International Gambling Studies*, *9*(1), 19–38. https://doi.org/10.1080/14459790802652209.

Delfabbro, P., & King, D. L. (2019). Challenges in the conceptualisation and measurement of gambling-related harm. *Journal of Gambling Studies*, *35*(3), 743–755. https://doi.org/10.1007/s10899-019-09844-1.

Deng, X., Lesch, T., & Clark, L. (2019). Applying data science to behavioral analysis of online gambling. *Current Addiction Reports*, *6*(3), 159–164. https://doi.org/10.1007/s40429-019-00269-9.

Dowling, N. A., Greenwood, C. J., Merkouris, S. S., Youssef, G. J., Browne, M., Rockloff, M., & Myers, P. (2021). The identification of Australian low-risk gambling limits: A comparison of gambling-related harm measures. *Journal of Behavioral Addictions*, *10*(1), 21–34. https://doi.org/10.1556/2006.2021.00012.

Drosatos, G., Arden-Close, E., Bolat, E., & Ali, R. (2020). Gambling data and modalities of interaction for responsible online gambling: A qualitative study. *Journal of Gambling Issues*, *44*. https://doi.org/10.4309/jgi.2020.44.8.

Drosatos, G., Nalbadis, F., Arden-Close, E., Baines, V., Bolat, E., Vuillier, L., … Bonello, M. (2018). Enabling responsible online gambling by real-time persuasive technologies. *Complex Systems Informatics and Modeling Quarterly*, *17*, 44–68. https://doi.org/10.7250/csimq.2018-17.03.

Ferris, J., & Wynne, H. (2001). *The Canadian problem gambling index*. Ottawa, ON: Canadian Centre on Substance Abuse.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal. Biometrische Zeitschrift*, *47*(4), 458–472. https://doi.org/10.1002/bimj.200410135.

Gainsbury, S. (2011). Player account-based gambling: Potentials for behaviour-based research methodologies. *International Gambling Studies*, *11*(2), 153–171. https://doi.org/10.1080/14459795.2011.571217.

Gainsbury, S. M. (2015). Online gambling addiction: The relationship between internet gambling and disordered gambling. *Current Addiction Reports*, *2*(2), 185–193. https://doi.org/10.1007/s40429-015-0057-8.

Garber, M. C., Nau, D. P., Erickson, S. R., Aikens, J. E., & Lawrence, J. B. (2004). The concordance of self-report with other measures of medication adherence: A summary of the literature. *Medical Care*, *42*(7), 649–652.

Gerstein, D., Volberg, R. A., Toce, M. T., Harwood, H., Johnson, R. A., Buie, T., … Engelman, L. (1999). *Gambling impact and behavior study: Report to the national gambling impact study commission*. Chicago: National Opinion Research Center.

Griffiths, M. (2003). Internet gambling: Issues, concerns, and recommendations. *CyberPsychology & Behavior*, *6*(6), 557–568. https://doi.org/10.1089/109493103322725333.

Haefeli, J., Lischer, S., & Schwarz, J. (2011). Early detection items and responsible gambling features for online gambling. *International Gambling Studies*, *11*(3), 273–288. https://doi.org/10.1080/14459795.2011.604643.

Harris, A., & Griffiths, M. D. (2017). A critical review of the harm-minimisation tools available for electronic gambling. *Journal of Gambling Studies*, *33*(1), 187–221. https://doi.org/10.1007/s10899-016-9624-8.

Heirene, R. M., Wang, A., & Gainsbury, S. M. (2021). Accuracy of self-reported gambling frequency and outcomes: Comparisons with account data. *Psychology of Addictive Behaviors*, *36*(4), 333–346. https://doi.org/10.1037/adb0000792.

Kainulainen, T. (2021). Does losing on a previous betting day predict how long it takes to return to the next session of online horse race betting? *Journal of Gambling Studies*, *37*, 609–622. https://doi.org/10.1007/s10899-020-09974-x.

Kairouz, S., Paradis, C., & Nadeau, L. (2012). Are online gamblers more at risk than offline gamblers? *Cyberpsychology, Behavior and Social Networking*, *15*(3), 175–180. https://doi.org/10.1089/cyber.2011.0260.

LaPlante, D. A., Nelson, S. E., & Gray, H. M. (2014). Breadth and depth involvement: Understanding Internet gambling involvement and its relationship to gambling problems. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *28*(2), 396–403. https://doi.org/10.1037/a0033810.

LaPlante, D. A., Nelson, S. E., LaBrie, R. A., & Shaffer, H. J. (2011). Disordered gambling, type of gambling and gambling involvement in the British Gambling Prevalence Survey 2007. *European Journal of Public Health*, *21*(4), 532–537. https://doi.org/10.1093/eurpub/ckp177.

Lesieur, H. R., & Blume, S. B. (1987). The South Oaks gambling screen (SOGS): A new instrument for the identification of pathological gamblers. *The American Journal of Psychiatry*, *144*(9), 1184–1188.

Lister, J. J., Nower, L., & Wohl, M. J. A. (2016). Gambling goals predict chasing behavior during slot machine play. *Addictive Behaviors*, *62*, 129–134. https://doi.org/10.1016/j.addbeh.2016.06.018.

Luquiens, A., Tanguy, M.-L., Benyamina, A., Lagadec, M., Aubin, H.-J., & Reynaud, M. (2016). Tracking online poker problem gamblers with player account-based gambling data only. *International Journal of Methods in Psychiatric Research*, *25*(4), 333–342. https://doi.org/10.1002/mpr.1510.

Macey, J., & Hamari, J. (2018). Investigating relationships between video gaming, spectating esports, and gambling. *Computers in Human Behavior*, *80*, 344–353. https://doi.org/10.1016/j.chb.2017.11.027.

McCormack, A., & Griffiths, M. D. (2013). A scoping study of the structural and situational characteristics of internet gambling. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, *3*(1), 29–49. https://doi.org/10.4018/ijcbpl.2013010104.

McCormack, A., Shorter, G. W., & Griffiths, M. D. (2014). An empirical study of gender differences in online gambling. *Journal of Gambling Studies*, *30*(1), 71–88. https://doi.org/10.1007/s10899-012-9341-x.

McCready, J., & Adlaf, E. (2006). Performance and enhancement of the Canadian problem gambling index (CPGI): Report and recommendations. https://doi.org/10.11575/PRISM/9894.

Miller, H. (2014). *Seeking help for gambling problems*. Victorian Responsible Gambling Foundation.

Miller, N. V., Currie, S. R., Hodgins, D. C., & Casey, D. (2013). Validation of the problem gambling severity index using confirmatory factor analysis and rasch modelling. *International Journal of Methods in Psychiatric Research*, *22*(3), 245–255. https://doi.org/10.1002/mpr.1392.

Nelson, S. E., LaPlante, D. A., Peller, A. J., Schumann, A., LaBrie, R. A., & Shaffer, H. J. (2008). Real limits in the virtual world: Self-limiting behavior of internet gamblers. *Journal of Gambling Studies*, *24*(4), 463–477.

Paluszynska, A., Biecek, P., Jiang, Y., & Jiang, M. Y. (2017). Package 'randomForestExplainer'. *Explaining and Visualizing Random Forests in Terms of Variable Importance*. https://cran.r-project.org/web/packages/randomForestExplainer/index.html.

Papineau, E., Lacroix, G., Sévigny, S., Biron, J.-F., Corneau-Tremblay, N., & Lemétayer, F. (2018). Assessing the differential impacts of online, mixed, and offline gambling. *International Gambling Studies*, *18*(1), 69–91. https://doi.org/10.1080/14459795.2017.1378362.

Percy, C., França, M., Dragičević, S., & Garcez, A. d'A. (2016). Predicting online gambling self-exclusion: An analysis of the performance of supervised machine learning models.

International Gambling Studies, *16*(2), 193–210. https://doi.org/10.1080/14459795.2016.1151913.

Perrot, B., Hardouin, J.-B., Costes, J.-M., Caillon, J., Grall-Bronnec, M., & Challet-Bouju, G. (2017). Study protocol for a transversal study to develop a screening model for excessive gambling behaviours on a representative sample of users of French authorised gambling websites. *BMJ Open*, *7*(5), e014600. https://doi.org/10.1136/bmjopen-2016-014600.

Perrot, B., Hardouin, J., Grall-Bronnec, M., & Challet-Bouju, G. (2018). Typology of online lotteries and scratch games gamblers" behaviours: A multilevel latent class cluster analysis applied to player account-based gambling data. *International Journal of Methods in Psychiatric Research*, *27*, e1746, Pubmed. https://doi.org/10.1002/mpr.1746.

Philander, K. S. (2014). Identifying high-risk online gamblers: A comparison of data mining procedures. *International Gambling Studies*, *14*(1), 53–63. https://doi.org/10.1080/14459795.2013.841721.

Powers, D. M. W. (2011). Evaluation, from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technology*, *2*(1), 37–63.

Price, A. (2022). Online gambling in the midst of COVID-19: A nexus of mental health concerns, substance use and financial stress. *International Journal of Mental Health and Addiction*, *20*(1), 362–379. https://doi.org/10.1007/s11469-020-00366-1.

Prochaska, J. O., & DiClemente, C. C. (2005). The transtheoretical approach. *Handbook of Psychotherapy Integration*, *2*, 147–171.

Rundle-Thiele, S. (2009). Bridging the gap between claimed and actual behaviour: The role of observational research. *Qualitative Market Research: An International Journal*, *12*(3), 295–306. https://doi.org/10.1108/13522750910963818.

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577.

Temcheff, C. E., Paskus, T. S., Potenza, M. N., & Derevensky, J. L. (2016). Which diagnostic criteria are most useful in discriminating between social gamblers and individuals with gambling problems? An examination of DSM-IV and DSM-5 criteria. *Journal of Gambling Studies*, *32*(3), 957–968. https://doi.org/10.1007/s10899-015-9591-5.

Toce-Gerstein, M., Gerstein, D. R., & Volberg, R. A. (2003). A hierarchy of gambling disorders in the community. *Addiction*, *98*(12), 1661–1672.

Tomei, A., Petrovic, G., & Simon, O. (2022). Offline and online gambling in a Swiss emerging-adult male population. *Journal of Gambling Studies*, 1–14. https://doi.org/10.1007/s10899-022-10106-w.

Ukhov, I., Bjurgert, J., Auer, M., & Griffiths, M. D. (2021). Online problem gambling: A comparison of casino players and sports bettors via predictive modeling using behavioral tracking data. *Journal of Gambling Studies*, *37*, 877–897. https://doi.org/10.1007/s10899-020-09964-z.

Welte, J. W., Barnes, G. M., Tidwell, M.-C. O., & Hoffman, J. H. (2009). The association of form of gambling with problem gambling among American youth. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *23*(1), 105–112. https://doi.org/10.1037/a0013536.

Wood, R. T., & Williams, R. J. (2011). A comparative profile of the Internet gambler: Demographic characteristics, gameplay patterns, and problem gambling status. *New Media & Society*, *13*(7), 1123–1141. https://doi.org/10.1177/1461444810397650.

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC press.