MODELO DE CLASIFICACIÓN PARA LA DETECCIÓN DE PERFILES DE RIESGO RESPECTO AL JUEGO

David Sanchez Corbella

IT Academy 2024

El presente informe detalla el desarrollo y los resultados obtenidos en el proyecto final del curso de Data Science, centrado en la creación de una herramienta de clasificación para detectar perfiles de riesgo relacionados con el juego. Además de describir el conjunto de datos utilizado, las variables analizadas y los métodos empleados, se ofrecen reflexiones sobre la importancia de la calidad de los datos, los desafíos al trabajar con conjuntos de datos preexistentes y la necesidad de supervisión de profesionales de la salud en este tipo de proyectos.

1. Objetivo y motivación

El principal propósito de este proyecto es la creación de una herramienta de clasificación destinada a la detección temprana de perfiles de riesgo en el juego dentro de la población. Se pretende alcanzar este objetivo mediante el análisis exhaustivo de las variables disponibles y la aplicación de modelos de aprendizaje automático, los cuales permitirán prever la probabilidad de que un individuo desarrolle problemas de juego en el futuro. Además de este objetivo central, se plantea evaluar la efectividad de diversas técnicas de procesamiento de datos y determinar las variables predictoras más relevantes en este contexto.

Como croupier con 24 años de experiencia en la industria del juego, comprendo profundamente la trascendencia de proyectos como este en la detección precoz de perfiles de riesgo asociados al juego. La prevención de problemas de juego es esencial para salvaguardar tanto a los individuos como a la sociedad en su conjunto, mitigando las consecuencias adversas derivadas del juego compulsivo. El desarrollo de herramientas de clasificación eficientes no solo tiene el potencial de identificar a aquellos en situación de riesgo, sino también de proporcionarles el apoyo necesario para abordar sus dificultades de manera oportuna.

Además, se considera que los resultados de este proyecto podrían ser de gran utilidad para la creación de una herramienta de autoevaluación destinada a los usuarios de juegos de azar. Una herramienta de este tipo permitiría a los individuos evaluar de manera autónoma su nivel de riesgo en relación con el juego, facilitando así la toma de conciencia y la búsqueda de ayuda profesional en caso necesario.

2. Presentación del Dataset

El conjunto de datos utilizado en este proyecto se deriva de un estudio exhaustivo sobre la prevalencia, comportamiento y características de los usuarios de juegos de azar en España, realizado por la Dirección General de Ordenación del Juego (DGOJ). La elección de este conjunto de datos se fundamenta en su significancia en el ámbito de la salud pública y la prevención de los problemas asociados al juego. Los datos fueron recolectados mediante encuestas y cuestionarios, y están disponibles de manera pública para su análisis en el siguiente enlace: Estudio de Prevalencia de la DGOJ.

Es importante destacar que, aunque la Ley 13/2011, de 27 de mayo, de regulación del juego, establece como objetivos ineludibles la protección de los menores y participantes de los juegos, así como la prevención de conductas adictivas, existen tan solo 3 estudios disponibles. El más reciente, del año 2023, constituye un valioso ejercicio analítico, pero lamentablemente, su conjunto de datos no está disponible debido a restricciones de protección de datos. Otro estudio, datado en 2017, se enfoca exclusivamente en la población con problemas de juego.Por último, el informe de 2015, el primero presentado por la DGOJ, sirvió como base para este proyecto. Este conjunto de datos está compuesto por 6816 filas y 240 columnas, abarcando así una amplia variedad de información relacionada con los usuarios de juegos de azar en España.

3. Características de las variables

Del dataset original, se realizó un filtrado para seleccionar únicamente a aquellas personas que respondieron afirmativamente a la pregunta: "¿Ha realizado algún tipo de juego con apuesta económica en el último año?", quedándonos con 4669 registros. Posteriormente, se llevó a cabo una selección de 38 columnas que abarcan una amplia variedad de información, desde aspectos demográficos hasta hábitos de juego y percepciones sobre la salud. Estas variables se distribuyen en diversas categorías, como características demográficas, hábitos de juego, percepciones y creencias relacionadas con el juego. Entre los factores seleccionados se incluyen la edad, el sexo, la situación laboral, la frecuencia de juego, la percepción del riesgo y la creencia en la suerte. Cada una de estas variables aporta información única que contribuye a la comprensión del comportamiento de los individuos en relación con el juego de azar.

En el anexo 1 que acompaña este informe se puede consultar una explicación detallada de cada una de las variables seleccionadas.

4. Variable Target

Una vez que hemos preparado el dataframe con el que vamos a trabajar, el siguiente paso es definir nuestra variable target para los modelos de clasificación supervisada. Para ello, recurrimos a una serie de preguntas presentes en la encuesta, las cuales pertenecen a 17 ítems utilizados en un estudio clínico de pacientes con problemas de juego (las cuales se pueden consultar en el anexo 2). En este estudio, se clasifica a los encuestados como personas sin riesgo (clase 0), personas con riesgo (clase 1) o personas con problemas (clase 2)

mediante una ponderación de estas preguntas y el resultado total de la suma de cada encuestado.

Para realizar esta clasificación inicial, utilizamos las respuestas a estas 17 preguntas, pero nos centramos únicamente en el último año. Además, utilizamos las mismas preguntas pero correspondientes a lo largo de la vida para identificar a personas que han tenido problemas en algún momento. Si una persona clasificada inicialmente como sin riesgo (clase 0) presenta problemas en estas respuestas de por vida, la reclasificamos como persona de riesgo (clase 1). Este tipo de modificaciones las hemos realizado también con otras 4 preguntas que fueron identificadas como marcadores por parte de un profesional en el ámbito de la salud mental, también presentes en el anexo 2.

Al final del proceso, obtuvimos un total de 4187 registros para perfiles de personas sin riesgo, 363 registros para personas con riesgo y 119 registros con problemas.

5. Limpieza de Datos y Análisis

A pesar de que el dataset se encuentra bien estructurado y compuesto únicamente por variables numéricas, su distribución es equitativa entre variables binarias (respuestas de sí o no) y ordinales (escalas del 0 al 5 o 7), lo que facilita la cuantificación de las respuestas y su análisis. Sin embargo, para mejorar su gestión, fue necesario realizar algunas transformaciones ligeras en ciertas columnas. Aunque no se detectaron valores nulos en el dataset, se encontraron múltiples casos de "No sabe no contesta", los cuales se abordaron de diversas formas, reconociendo un sesgo en el análisis pero considerándolo necesario para su posterior procesamiento.

Se llevó a cabo un análisis exploratorio de datos con el objetivo de examinar las distribuciones de las variables y buscar posibles relaciones entre ellas. Este análisis se basó en un gráfico de correlaciones, así como en la utilización de *boxplots* y *kdeplots* para proporcionar una visualización clara de estas estadísticas.

6. Modelos y resultados

En la creación de modelos, nuestro principal objetivo era lograr altos números en el Recall de la clase 1, con el fin de diagnosticar correctamente los perfiles de riesgo. No se le dio tanta importancia a los falsos positivos que pudieran surgir entre los grupos de personas con riesgos y problemas, ya que estudios indican que cerca del 90% de estos últimos tienen conciencia de sus problemas por sí mismos.

A partir de aquí, exploramos cuál podría ser nuestro mejor modelo entre KNN, XGBoost, SVC y Random Forest. Aunque los resultados para la clase 0 eran satisfactorios, para el resto de las clases eran deficientes. Esto se debió a la notable descompensación entre las clases de la variable objetivo.

Para intentar abordar este problema, recurrimos a la ponderación de clases utilizando el modelo SVC, el cual nos permitió asignar mayores pesos a nuestras clases más relevantes. Sin embargo, los resultados seguían siendo decepcionantes.

Finalmente, implementamos una técnica de sobremuestreo de clases minoritarias, como SMOTE, lo que nos llevó a obtener números impresionantes, con un promedio de Recall cercano al 0.97. No obstante, al realizar la Validación Cruzada con nuestro conjunto de datos sin sobremuestrear utilizando Random Forest, que fue nuestro mejor modelo, volvimos a obtener resultados muy desalentadores en la clasificación de la clase 1.

7. Conclusiones

Conjuntos de datos de alta calidad: La calidad de los datos es fundamental en cualquier proyecto de análisis o modelado, y esto es especialmente relevante en el campo de la detección y prevención de problemas de juego. Los conjuntos de datos deben ser precisos, completos y representativos de la población objetivo. Esto implica la necesidad de recopilar datos de manera adecuada, utilizando metodologías rigurosas y validadas. Además, es importante que los datos estén actualizados y sean relevantes para los objetivos del proyecto.

Importancia de modelar todo el proyecto y desafíos al trabajar con datasets preexistentes: Es crucial modelar integralmente un proyecto desde su concepción hasta la implementación de modelos para garantizar la efectividad y la relevancia de los resultados obtenidos. Modelar un proyecto en su totalidad permite una comprensión más profunda de los datos, los objetivos y las posibles limitaciones, lo que a su vez facilita la creación de modelos más precisos y útiles. Sin embargo, trabajar con datasets preexistentes diseñados para análisis de datos puede plantear desafíos significativos. Estos conjuntos de datos pueden no estar optimizados para la modelización predictiva y pueden carecer de variables clave o presentar desequilibrios en las clases objetivo. Por lo tanto, adaptar estos datasets para la creación de modelos puede requerir una cuidadosa evaluación, preprocesamiento y selección de características para garantizar la calidad y la eficacia de los modelos resultantes.

Supervisión de profesionales de la salud: La participación de profesionales del ámbito de la salud, como psicólogos clínicos o trabajadores sociales, es crucial para garantizar que los enfoques y herramientas utilizadas en el proyecto sean éticos y efectivos desde el punto de vista clínico. Estos profesionales pueden aportar su experiencia en la evaluación y el tratamiento de los problemas de juego, así como en la atención de las necesidades de las personas afectadas y sus familias. Además, aprovechando la colaboración interdisciplinaria, la supervisión de profesionales de la salud permite integrar perspectivas especializadas, como la psicología y la salud pública, para abordar el juego problemático desde un enfoque integral y multidimensional. Esto garantiza que se consideren adecuadamente los aspectos psicológicos, sociales y de salud pública relacionados con el problema del juego problemático, mejorando así la efectividad de las estrategias de intervención y prevención.

Seguimiento de estudios y continua investigación: Dado que el campo del juego problemático está en constante evolución, es importante realizar un seguimiento de los estudios realizados y estar al tanto de las últimas investigaciones y avances en el campo. Esto permite incorporar nuevas evidencias y enfoques en los proyectos futuros, así como mejorar las herramientas y técnicas disponibles para la detección y prevención de problemas de juego. La investigación continua también es fundamental para identificar nuevas tendencias y desafíos en el campo y adaptar las estrategias de intervención en consecuencia.

Anexo 1. Definición de las Variables

edad: Edad

sexo: Sexo:
1 Hombre - 2 Mujer

estudios: Nivel de educación alcanzado
1 Sin estudios
2 Primarios
3 Secundarios
4 Superiores

situac_laboral: Situación laboral actual
1 Trabajando
2 Desempleado
3 Estudiante
4 Jubilado o prejubilado
5 Incapacitado para trabajar
6 Labores de hogar

estado_civil: Estado civil
1 Soltero
2 Casado, viviendo en pareja
3 Separado / Divorciado
4 Viudo

val_salud: Valoración del estado de salud:
1 (muy mala) - 5 (muy buena)

alcohol: Frecuencia de consumo de alcohol:
0 (nunca) - 6 (diariamente varias veces)

fuma: Frecuencia de consumo de tabaco:
0 (nunca) - 6 (más de dos cajetillas)

problema_psico: Existencia de algún problema psicológico:
1 Sí - 2 No

ayuda_ansie: Recurso a ayuda profesional para controlar episodios de ansiedad:
1 Sí - 2 No

impulsivo: Nivel de impulsividad:
1 (nada impulsiva) - 5 (muy impulsiva)

satisf_vida: Grado de satisfacción con la vida:
0 (nada satisfecho) - 10 (muy satisfecho)

online: Experiencia previa en juegos de azar en línea:
1 Sí - 2 No

horas: Horas dedicadas semanalmente a juegos de azar en línea:
1 (menos de 1 hora) - 6 (más de 20 horas)

val_horas: Valoración del tiempo dedicado a juegos de azar en línea:
1 (poco) - 5 (excesivo)

gasto: Gasto promedio mensual en juegos de azar
0 Menos de 10€
1 Entre 10 y 50 €
2 Entre 50,01 y 100 €
3 Entre 100,01 y 300€
4 Más de 300

val_gasto: Valoración del gasto en juegos de azar:
1 (poco) - 5 (excesivo)

entorno: Experiencia de problemas con el juego en el entorno familiar o social:
1 Sí - 2 No

pref_premios: Preferencia por premios inmediatos o prolongados en juegos de azar
1 Premios inmediatos
2 Juegos prolongados con mayores recompensas

planificacion: Nivel de planificación del tiempo dedicado a juegos de azar:
1 Sí - 2 No

control: Capacidad de detenerse al jugar según propia decisión:
1 Sí - 2 No

espera: Molestias experimentadas por tiempos de espera en juegos de azar:
1 Sí - 2 No

frec_loteria: Frecuencia de participación en loterías nacionales, primitivas y ONCE:
1 (nunca) - 7 (todos los días)

frec_slots: Frecuencia de uso de máquinas tragamonedas y recreativas:
1 (nunca) - 7 (todos los días)

frec_apuestas: Frecuencia de participación en otras formas de apuestas como casinos, bingo y apuestas deportivas:
1 (nunca) - 7 (todos los días)

online_loteria: Experiencia previa en loterías nacionales, primitivas y ONCE en línea:
1 (nunca) - 7 (todos los días)

online_slots: Experiencia previa en máquinas tragamonedas y recreativas en línea:
1 (nunca) - 7 (todos los días)

online_apuestas: Experiencia previa en otras formas de apuestas en línea como casinos, bingo y apuestas deportivas:
1 (nunca) - 7 (todos los días)

habilidad: Creencia en la influencia de la habilidad en juegos de azar:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

probab_repetir: Creencia en la probabilidad de repetir un premio importante:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

aprove_racha: Creencia en la importancia de aprovechar una racha de suerte:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

persistencia: Creencia en la necesidad de persistir en el juego para obtener ganancias:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

cercania_ganar: Creencia en la cercanía de la victoria como señal de éxito próximo:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

creencia_suerte: Creencia en la existencia de la suerte en juegos de azar:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

predic_premios: Creencia en la posibilidad de predecir los premios en juegos de azar:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

sensa_especial: Experiencia de sensaciones especiales que anticipan la victoria en juegos de azar:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

recuperar: Creencia en la capacidad de recuperar pérdidas al jugar más tiempo:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

incentivo: Creencia en la existencia de premios de incentivo para continuar jugando:
1 (totalmente en desacuerdo) - 5 (totalmente de acuerdo)

Anexo 2. Preguntas para variable target

P14A1	¿Ha tenido períodos de 2 o más semanas en las que pasase una gran cantidad de tiempo pensando en sus experiencias con el juego o planificando detalladamente futuros episodios de juego o de apuestas?
P14A2	¿Ha tenido períodos de 2 o más semanas en las que pasase mucho tiempo pensando en cómo conseguir dinero para jugar?
P14A3	¿Ha tenido períodos de 2 o más semanas en los que necesitaba jugar con cantidades de dinero cada vez mayores, o apuestas mayores que antes, para conseguir la misma excitación?
P14A4	¿Ha intentado alguna vez dejar, reducir o controlar su juego?
P14A5	En una o más de estas ocasiones de intento de dejar, reducir o controlar su juego, ¿se sintió intranquilo o irritable?
P14A6	¿Alguna vez ha intentado dejar, reducir o controlar su conducta de juego sin poder conseguirlo?
P14A7	En el caso de que fuese así, ¿ha sucedido 3 o más veces?
P14A8	¿Ha jugado alguna vez como una forma de escapar de los problemas personales?
P14A9	¿Ha jugado alguna vez para aliviar sentimientos desagradables como culpabilidad, ansiedad, indefensión o depresión?
P14A10	¿Ha tenido alguna vez un período en el cual si perdía dinero en el juego volvía otro día para recuperarlo?
P14A11	¿Ha mentido alguna vez a su familia, amigos o a otros sobre cuánto juega o cuánto dinero perdía en el juego?
P14A12	Si es así, ¿esto ha sucedido 3 o más veces?
P14A13	¿Ha extendido alguna vez un cheque sin fondos o cogido dinero que no era suyo de familiares u otra persona para gastar en el juego?
P14A14	¿Le ha causado alguna vez el juego problemas graves o repetidos en su relación con algún familiar o amigo?
P14A15	¿Le ha producido el juego algún problema con los estudios, como por ejemplo perder clases o días de escuela o suspender algún curso?
P14A16	¿Le ha causado el juego la pérdida de un trabajo, tener problemas en el trabajo o no poder aprovechar una oportunidad profesional importante?

P14A17 ¿Ha necesitado alguna vez pedir dinero prestado a un familiar, o a otra persona, para poder salir de una situación económica desesperada causada en gran parte por el juego?

P15A ¿Ha tenido usted alguna vez problemas de juego excesivo o dependencia con algún juego de azar a lo largo de su vida?

P15B ¿Y en el último año?

P16A ¿Su familia o amigos le dicen que tiene dependencia con algún juego de azar o que juega excesivamente?

P16B ¿Diría vd. que tienen razón en su opinión de que juega excesivamente?

P43 ¿Ha recurrido a ayuda profesional para controlar alguna adicción?