UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

Evaluación de la Disfunción Ejecutiva en Trastornos Cerebrales Mediante Tecnologías de Videojuegos

Diego Hernán Montenegro Ducaud

TESIS PARA OPTAR AL GRADO DE MAGISTER EN INFORMÁTICA MÉDICA.

Director de Tesis: Prof. Dr. Mauricio Cerda Villablanca
Codirectora de Tesis: Prof. Dr. Andrea Slachevsky Chonchol
Codirector de Tesis: Prof. Dr. David Martínez-Pernía

2025

Agradecimientos

Llegar hasta aquí fue más parecido a superar un nivel difícil que a seguir un tutorial paso a paso. Por eso quiero dejar constancia, a modo de *créditos finales*, de quienes hicieron posible este juego llamado tesis.

En principio, agradecer a mi hermana Natalia, por el apoyo constante (en comida y café principalmente) y el ánimo otorgado por todo este tiempo. A mi pareja, Javiera, por su paciencia, entusiasmo y por sobre todo, por las risas y mensajes para poder seguir adelante en esto: *hasta la ecuación más compleja se resuelve paso a paso*.

A mi madre y sobrinos, que, aunque no entendían del todo por qué pasaba noches peleando con datos y código, me miraban con orgullo: este esfuerzo también es para ustedes, como prueba de que *sí se puede* y para que construyan algo todavía más grande.

A mis amigos que supieron alentarme, darme un espacio de descanso y conversación poco seria cuando más se necesita. Mención especial a Seta, por animarme a sustentar esta tesis y terminarla antes de que él lo haga (nunca).

En el terreno académico, expreso mi más sincero agradecimiento al profesor Mauricio Cerda, por su guía constante, su exigencia y, sobre todo, la confianza depositada en mi trabajo. Desde el inicio me brindó oportunidades y me empujó fuera de mi zona de confort para seguir creciendo.

También a la Dra. Andrea Slachevsky y al Dr. David Martínez-Pernía, co-directores de este proyecto, por compartir su experiencia, orientarme y aportar una mirada interdisciplinaria que enriqueció cada etapa del estudio.

Agradezco al equipo FONDEF IDeA I+D 2022 #ID22I10251 por la financiación y el respaldo logístico que hicieron viable el desarrollo de Neuronat, así como a la Mutual de Seguridad y al Hospital del Salvador por su colaboración en el reclutamiento y la evaluación de participantes.

¡*Press Start* para lo que viene!

Índice

Índice de Tablas

Índice de Figuras

Resumen

Las funciones ejecutivas (FE) permiten planificar, organizar, resolver problemas y adaptarse a situaciones cambiantes. Su alteración puede observarse en personas con traumatismo encéfalo craneano (TEC) o secuelas post-COVID-19, comprometiendo su funcionalidad diaria. Las pruebas neuropsicológicas tradicionales, aunque efectivas para detectar disfunción ejecutiva (DE), presentan baja validez ecológica y limitaciones para identificar casos leves o intermedios.

Frente a este desafío, la presente tesis explora el uso de Neuronat, un Serious Game que simula la gestión de un restaurante, como herramienta para evaluar DE. En este entorno, los participantes deben realizar múltiples tareas, tomar decisiones y resolver dilemas, mientras se recopilan métricas de su desempeño. Se propone que estos datos pueden alimentar modelos de aprendizaje automático (*Machine Learning*) para predecir la presencia de DE con base en el juicio clínico experto.

Se desarrolló un pipeline que incluyó el procesamiento de archivos JSON, extracción de más de 80 variables por sesión, reducción de dimensionalidad y entrenamiento de clasificadores supervisados (*Random Forest, Logistic Regression, XGBoost*, entre otros), validados mediante validación cruzada anidada. Además, se compararon diferentes conjuntos de variables (demográficas, neuropsicológicas, métricas de juego, métricas de grafos).

Los modelos entrenados exclusivamente con datos derivados del juego alcanzaron métricas destacadas, con F1-score promedio por sobre 0.70 y AUC superiores a 0.78. Este desempeño es comparable estadísticamente al de modelos basados en evaluaciones clínicas tradicionales, lo que respalda la utilidad de Neuronat como instrumento complementario. Además, su alta usabilidad y bajo sesgo de administración aseguran que las diferencias observadas reflejan la capacidad cognitiva y no dificultades técnicas.

En conclusión, Neuronat representa una herramienta válida, automatizable y con alta validez ecológica para apoyar el tamizaje y monitoreo clínico de la disfunción ejecutiva. Este trabajo aporta evidencia empírica para avanzar en la integración de *Serious Games* como instrumentos complementarios en evaluación neuropsicológica.

Abstract

Executive functions (EF) enable individuals to plan, organize, solve problems, and adapt to changing situations. Their impairment can be observed in individuals with traumatic brain injury (TBI) or post-COVID-19 cognitive sequelae, compromising daily functioning. Although traditional neuropsychological tests are effective in detecting executive dysfunction (ED), they often lack ecological validity and show limitations in identifying mild or moderate cases.

To address this challenge, this thesis explores the use of *Neuronat*, a serious game that simulates restaurant management, as a tool for evaluating ED. In this environment, participants are required to perform multiple tasks, make decisions, and solve dilemmas, while metrics on their performance are recorded. It is proposed that these data can be used to train machine learning models to predict the presence of ED based on expert clinical judgment.

A computational pipeline was developed, including the processing of JSON files, extraction of more than 80 variables per session, dimensionality reduction, and training of supervised classifiers (Random Forest, Logistic Regression, XGBoost, among others), validated through nested cross-validation. In addition, different sets of variables were compared (demographic, neuropsychological, gameplay-based, and graph-based metrics).

Models trained exclusively on gameplay-derived data achieved outstanding results, with average F1-scores above 0.70 and AUC values exceeding 0.78. This performance was statistically comparable to that of models based on traditional clinical evaluations, supporting the usefulness of *Neuronat* as a complementary instrument. Furthermore, its high usability and low risk of administration bias ensure that observed differences reflect cognitive capacity rather than technical limitations.

In conclusion, *Neuronat* represents a valid, automatable, and ecologically sound tool to support the screening and clinical monitoring of executive dysfunction. This study provides empirical evidence to advance the integration of serious games as complementary instruments in neuropsychological assessment.

1. Introducción

1.1 Antecedentes

Funciones Ejecutivas

Las funciones ejecutivas (FE) son un conjunto de procesos cognitivos de alto nivel que facilitan la gestión y regulación de conducta y pensamiento. Diversos autores indican que las funciones ejecutivas pueden conceptualizarse como un "concepto paraguas" debido a su naturaleza multidimensional, abarcando desde autocontrol hasta flexibilidad cognitiva y planificación (Goldstein et al., 2014; Burgess et al., 2006), esenciales para la salud física y mental, el éxito en la vida, y el desarrollo cognitivo, social y psicológico (Diamond, 2020, 2013).

Aunque tradicionalmente se ha considerado a las FE como una habilidad general de alto orden, se propone que comprenden tres procesos centrales: inhibición, memoria de trabajo y flexibilidad cognitiva (e.g., Lehto et al., 2003, Miyake et al., 2000; McAlister et al., 2016). A partir de estas tres, se construyen funciones ejecutivas de orden superior como resolución de problemas, razonamiento y planificación. (Collins & Koechlin, 2012, Lunt et al., 2012).

Las FE comprenden una amplia gama de subdominios cognitivos y competencias conductuales, como como la flexibilidad cognitiva, la capacidad de pasar de una tarea a otra (*shifting*), inhibición de respuestas automáticas no adecuadas al contexto, la capacidad de mantener en mente la información necesaria para llevar a cabo las acciones en curso, la capacidad de modificar el comportamiento para adaptarse al contexto, la multitarea, la resolución de problemas, la resistencia a las interferencias, la autoconciencia, la metacognición y la capacidad de abordar la novedad (Burgess et al., 2000; Chan et al., 2008), entre otras.

La disfunción ejecutiva (DE), definida como el deterioro de las FE, se describe como un factor principal asociado con el deterioro funcional en trastornos neuropsiquiátricos y neurológicos, manifestándose como problemas en el manejo de rutina diaria o la labilidad emocional (Martínez-Pernía et al., 2023). Estas funciones pueden verse afectadas por diversas enfermedades y trastornos neurológicos, como tumores

cerebrales (Robinson et al., 2015), enfermedades cerebrovasculares (Hua et al., 2014), la enfermedad de Parkinson, traumatismo encéfalo craneano (TEC) (Karr et al., 2014) esclerosis múltiple, síndrome de Tourette, trastorno por déficit de atención e hiperactividad (TDAH) (Cummings, 1993; Pennington & Ozonoff, 1996), pero también en trastornos psiquiátricos, por ejemplo, esquizofrenia (Xu et al., 2014) o trastorno obsesivo-compulsivo (Nakao et al., 2014), entre otros. En esta tesis los pacientes con disfunción ejecutiva pertenecerán a dos grupos: Pacientes TEC y pacientes COVID.

Después de un TEC, las personas pueden experimentar problemas con la atención y concentración, capacidad para iniciar y detener acciones, toma de decisiones, planificación y organización de tareas y flexibilidad cognitiva (McDonald et al., 2002). Estas dificultades pueden provocar retos en tareas cotidianas y en el lugar de trabajo, y pueden generar cambios en el comportamiento y la personalidad, afectando las relaciones interpersonales (Stuss & Levine, 2002).

La infección por COVID-19 ha sido asociada con una variedad de complicaciones neuropsicológicas. Entre el 44% y el 53% de los pacientes que han padecido COVID-19 reportan alteraciones cognitivas, con un impacto significativo en las funciones ejecutivas (Varatharaj et al., 2020). Éstas incluyen dificultades en atención, planificación, flexibilidad cognitiva, memoria de trabajo y toma de decisiones, pudiendo ser el resultado de la gravedad de la infección, la respuesta inflamatoria sistémica, y posibles efectos del virus en el sistema nervioso central (Helms et al., 2020). Al igual que en el TEC, estos pacientes pueden enfrentar retos en gestión de actividades diarias y laborales, y cambios en conducta y personalidad.

Una evaluación rigurosa de las funciones ejecutivas se ha mostrado como un sólido predictor del desempeño y la participación funcional tras eventos neurológicos (Adamit et al., 2015; Romero-Ayuso et al., 2019). Esta idea predictiva se mantiene a lo largo de todo el ciclo vital, sin verse limitado por la edad (Diamond, 2013) y persiste incluso después de controlar la gravedad inicial de la lesión cerebral subyacente (Kinnunen et al., 2011), por lo cuál es importante detectar de forma temprana una disfunción ejecutiva.

Evaluación de las Disfunciones Ejecutivas

La evaluación de las funciones ejecutivas es hoy en día es principalmente tradicional, es decir, se usan medios como test de lápiz y papel y/o entrevistas ya sea al paciente como también a personas cercanas al paciente para evaluar la presencia de una DE. Los diferentes tipos de evaluación se presentan en la Tabla 1.

Pruebas en papel y lápiz	Observación Naturalista	TIC
Pruebas basadas dominios cognitivos específicos:	Multiple Errands Test (MET) (Shallice & Burgess, 1991)	***Evaluación neuropsicológica por computador:***
Rey Complex Figure Test (Osterrieth, 1944)	Executive Function Route-finding Task (EFRT) (Boyd & Sautter, 1993)	Tower of Hanoi (Mataix-Cols & Bartres-Faz, 2002)
Rivermead Behavioral Memory Test (Wilson et al., 1985)	Executive Function Performance Test (EFPT) (Baum et al., 2008; Weiner et al., 2012)	The Category Test (Choca & Morris, 1992)
Test of Everyday Attention (Manly et al., 2001)	Functional Independence Measure (FIM) (Keitll et al., 1987).	Raven Progressive Matrices Test (Williams & McCord, 2006)
Wisconsin Card Sorting Test (WCST) (Axelrod et al., 1996)		CANTAB (Sahakian & Owen, 1992)
Trail Making Test (TMT) (Reitan, 1958)		***Realidad Virtual:***
Baterías neuropsicológicas:		V-Store (Lo Priore et al., 2003)
Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1939)		Virtual Action Planning-Supermarket (VAP-S) (Klinger et al., 2004)
Luria-Nebraska Battery (Golden et al., 1985)		Virtual Mall (Rand et al., 2005)
Cuestionarios:		Virtual MET (Rand et al., 2009)
Behavioral Assessment of the Dysexecutive Syndrome (Wilson et al., 1996).		***Serious Games:***
Behavioral Dysexecutive Syndrome Inventory (BDSI) (Godefroy et al., 2010)		Virtual Reality Grocery (V-Store) (Levy et al., 2019)
The Neuropsychiatric Inventory (NPI) (Cummings et al., 1994)		The multitasking in the city test (MCT) (Jovanovski, Zakzanis, Campbell, et al., 2012; Jovanovski, Zakzanis, Ruttan, et al., 2012)
		Non-immersive Virtual Coffee task (NI-VCT) (Besnard et al., 2016).

Tabla 1. Pruebas para detectar Disfunción Ejecutiva (Martínez-Pernía et al., 2023)

Las actividades cotidianas implican diversas habilidades cognitivas que pueden verse comprometidas en presencia de una disfunción ejecutiva (DE), por lo tanto, es fundamental desarrollar métodos de evaluación que no solo midan las capacidades cognitivas en entornos controlados, sino que también sean capaces de predecir el desempeño en contextos cotidianos reales, de ahí cobra relevancia el concepto de validez ecológica, el cual se refiere a la capacidad de extrapolar o generalizar los resultados obtenidos en un laboratorio o entorno controlado al comportamiento natural en el mundo real (Schmuckler, 2001). Esto es crucial para obtener una comprensión completa de las capacidades ejecutivas de los pacientes y para diseñar intervenciones más efectivas y personalizadas.

I. Pruebas de Lápiz y Papel

Las pruebas de lápiz y papel han sido históricamente la base de evaluación neuropsicológica de las FE. Herramientas como la Prueba de Clasificación de Cartas de Wisconsin (WCST) y la Prueba de Stroop han proporcionado información valiosa sobre capacidades cognitivas de los individuos (Lezak et al., 2012). A pesar de su amplio uso, enfrentan limitaciones significativas, como la falta de validez ecológica, al no replicar fielmente demandas cognitivas complejas de la vida diaria (Chaytor et al., 2006).

Por otro lado, la necesidad de un evaluador experto y los desafíos de accesibilidad para ciertos grupos demográficos limitan su aplicabilidad (Brearly et al., 2017; Valladares-Rodríguez et al., 2016).

Una proporción de la variabilidad en los resultados funcionales no se explica por los rendimientos en estas tareas cognitivas, lo que muestra que presentan una capacidad limitada para predecir el funcionamiento diario (McAlister et al., 2016). Por ejemplo, algunas personas con deterioro cognitivo significativo mantienen un buen nivel de funcionamiento en el mundo real, mientras que otras con menos deterioro cognitivo están más discapacitadas funcionalmente.

II. Evaluaciones basadas en observaciones naturalistas.

Se han desarrollado métodos basados en la observación naturalista para evaluar FE mientras el paciente realiza actividades de la vida diaria. Algunos ejemplos incluyen el *Multiple Errands Test* (MET) (Shallice & Burgess, 1991), el *Executive Function Route-finding Task* (EFRT) (Boyd & Sautter, 1993), el *Executive Function Performance Test* (EFPT) (Baum et al., 2008; Weiner et al., 2012) y el *Functional Independence Measure* (FIM) (Keitll et al., 1987). Estos métodos suelen implementarse en entornos simulados diseñados para replicar situaciones reales y se centran en la actividad y participación del paciente, más que en su nivel de discapacidad, como se puede observar en la Figura 1.

Las evaluaciones en contextos reales han mostrado ser más precisas que las realizadas en laboratorio (Burgess et al., 2006; Rand et al., 2009). Sin embargo, presentan limitaciones como el alto consumo de tiempo y la dificultad para obtener medidas estandarizadas en diferentes centros de administración (Chevignard et al., 2000; Fortin et al., 2003).

Figura 1. Evaluación con el Executive Function Performance Test (EFPT) La imagen muestra a un terapeuta ocupacional supervisando a un paciente mientras organiza medicamentos, una de las tareas del EFPT. Fuente: Mizzou College of Health Sciences, "Scoring the Executive Function Performance Test" https://youtu.be/SHM4MQLowZY?si=vJyJXN9qL--JTQWx

III. Realidad Virtual

La realidad virtual (RV) ha emergido como herramienta innovadora en evaluación de las FE. A través de entornos computarizados, como se puede observar en la Figura 2, la RV ofrece experiencias inmersivas que simulan situaciones de la vida real, con una validez ecológica superior a las pruebas tradicionales (Riva et al., 2007; Parsons, 2015), ya que esta tecnología permite la manipulación de elementos del entorno para estudiar respuestas específicas de los participantes, algo que es difícil de lograr con métodos tradicionales (Gorini et al., 2011; Diemer et al., 2015) A pesar de todo, la RV enfrenta limitaciones como malestar en algunos usuarios y requisitos de habilidades técnicas y recursos financieros (Bohil et al., 2011). Estudios han demostrado que tareas realizadas en entornos de RV están significativamente asociadas con evaluaciones tradicionales, proporcionando así una herramienta más precisa y representativa para evaluar las FE (Armstrong et al., 2013).

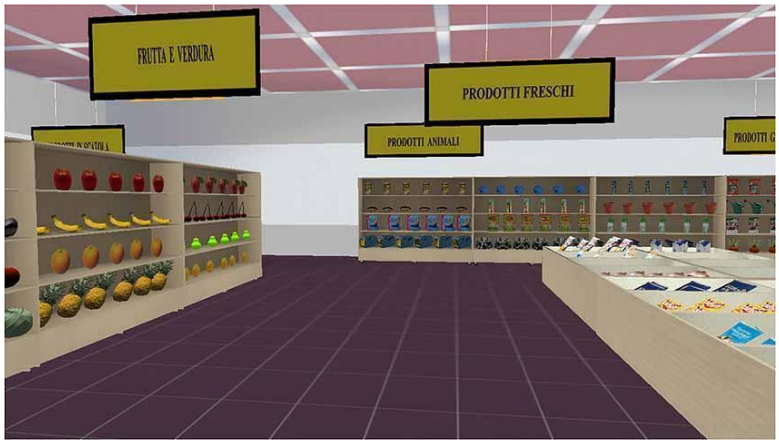


Figura 2. Captura de pantalla de VMET (Serino et al., 2014). herramienta basada en realidad virtual, para evaluar funciones ejecutivas en pacientes con Parkinson, realizando tareas cotidianas, como comprar artículos en una tienda, enviar cartas y recoger información en diferentes lugares.

IV. Serious Games

Los Serious Games (SG) representan una alternativa en la evaluación y manejo de las disfunciones ejecutivas a través de las Tecnologías de la Información y la Comunicación (TIC). Definidos como juegos digitales especializados con propósitos más allá del entretenimiento (Michael & Chen, 2005; Robert et al., 2014).

Estos juegos han sido aplicados en poblaciones diversas, incluyendo niños, adultos y adultos mayores, tanto en poblaciones sanas como en aquellas con deterioro cognitivo, para evaluar la disfunción ejecutiva (Jansari et al., 2014; Neto et al., 2018; Valladares-Rodriguez et al., 2019; Van de Weijer-Bergsma et al., 2015; Wang et al., 2022). Ejemplos destacados incluyen el entorno virtual *V-mart* de Levy et al. (2019), donde los participantes realizan tareas de manejo del tiempo, presupuesto y planificación en una tienda de comestibles; el *Multitasking in the City Test* (MCT) de Jovanovski, Zakzanis, Campbell, et al., (2012), que involucra la realización de tareas mientras se navega por una ciudad virtual; y el *Non-immersive Virtual Coffee Task* (NI-VCT) de Besnard et al. (2016), donde los participantes deben preparar una taza de café utilizando una máquina de café virtual, como se ve en la Figura 3:

Figura 3. Captura de pantalla de la cocina virtual del NI-VCT (Besnard et al., 2016)

Una de las ventajas de los SG es que recopilan una gran cantidad de datos detallados sobre las acciones y decisiones del jugador en tiempo real, registrando información que en una evaluación de lápiz y papel es más complicado de capturar. Esto permite un análisis más profundo y preciso de las capacidades cognitivas y comportamentales de los participantes, por ejemplo tiempos de respuestas, precisión, errores o estrategias utilizadas, pueden almacenarse en datos de guardado para un posterior análisis.

En comparación con las pruebas neuropsicológicas tradicionales, los Serious Games presentan ventajas significativas en cuanto a validez ecológica, ya que simulan situaciones cercanas a la vida diaria del paciente, y permiten evaluar funciones ejecutivas en contextos más naturales y dinámicos.

A diferencia de estas pruebas convencionales, que suelen ofrecer puntuaciones globales obtenidas en sesiones únicas y en ambientes clínicos controlados, los SG permiten registrar variables discretas y continuas en tiempo real, tales como errores, estrategias, tiempos de espera, decisiones éticas, rutas de navegación y eficiencia en múltiples tareas paralelas. Esta riqueza de datos abre la puerta a evaluaciones más precisas, personalizadas y sensibles al cambio, lo que resulta particularmente útil para el seguimiento longitudinal y la intervención cognitiva.

Además, los Serious Games pueden implementarse en entornos accesibles y de bajo costo, facilitando su integración en contextos clínicos o comunitarios (Wattanasoontorn et al., 2013). Su carácter interactivo y gamificado favorece la adherencia de los usuarios y permite recolectar datos con alta granularidad sin comprometer la experiencia del participante. En este sentido, los SG no pretenden necesariamente reemplazar las herramientas neuropsicológicas tradicionales, sino ampliar el espectro de observación clínica, facilitando evaluaciones más naturales, sensibles al contexto y centradas en el desempeño cotidiano del individuo (Valladares-Rodríguez et al., 2016).

Diversos autores han contrastado las evaluaciones neuropsicológicas tradicionales con los enfoques emergentes basados en tecnologías interactivas como los Serious Games. La siguiente tabla resume algunas diferencias clave, agrupadas según criterios de administración, aplicabilidad, motivación y análisis de datos, basadas en estudios comparativos como los de Tong et al. (2016) y Tso et al. (2015). Esta comparación permite visualizar cómo los SG introducen ventajas operativas relevantes para entornos clínicos y de investigación, aunque todavía enfrentan desafíos en cuanto a su validación estandarizada, como se puede apreciar en la Tabla 2:

Dimensión de Evaluación	Evaluación Tradicional	Serious Game
Método de administración	Evaluador capacitado	Autoaplicación
Sesgo de administración	✓	✗
Equipo de aplicación	Lápiz y papel	Tablet/Computadora
Repetibilidad / variabilidad	Limitada al tiempo y formularios	Limitada a la aleatoriedad
Motivación / Entretenimiento	Baja	✓
Validación	Disponible según prueba	Aún por validar
Análisis de datos	Según tabulación	Entrega inmediata
Múltiples lenguajes	Según validación del lugar	En la configuración
Funciones de pausa	Entre sub-pruebas	Disponible durante el juego

Tabla 2. Comparación entre Evaluación Tradicional y Serious Game Adaptado de Tong et al., 2016 y Tso et al., 2015

Aplicación de Machine Learning en el Diagnóstico de Enfermedades

En el campo de la medicina, se han realizado estudios incorporando modelos computacionales de *Machine Learning* (ML) en la evaluación y diagnóstico de distintas patologías. Por ejemplo, se han desarrollado modelos de ML que, basados en imágenes de resonancias magnéticas, permiten diagnosticar etapas tempranas de la enfermedad de Alzheimer, diferenciando entre Alzheimer, discapacidad cognitiva leve y deterioro natural por envejecimiento (Khedher, Ramírez, Górriz, Brahim, & Segovia, 2020). Otro ejemplo es el uso de herramientas de ML para diagnosticar ansiedad y depresión en niños mediante datos generados a partir de tests específicos (McGinnis et al., 2019).

Los sistemas de aprendizaje automático en medicina, especialmente en el diagnóstico de enfermedades, según Kononenko (2001) los algoritmos de ML pueden mejorar la precisión diagnóstica mediante el análisis de grandes volúmenes de datos clínicos y la identificación de patrones complejos que podrían pasar desapercibidos para los médicos. Estos sistemas pueden integrarse con métodos estadísticos y expertos clínicos para crear modelos híbridos.

Neuronat: Un Serious Game para Evaluar Disfunción Ejecutiva

Neuronat es un Serious Game diseñado para el apoyo en el diagnóstico y la evaluación de disfunción ejecutiva, particularmente en contexto de evaluaciones neuropsicológicas. Desarrollado como herramienta virtual no inmersiva, se administra bajo supervisión de un profesional de la salud.

Neuronat sitúa a los jugadores en un restaurante italiano, asumiendo el rol de un mesero, como se observa en la Figura 4. Comienza con una fase tutorial, en la cual un personaje (el chef) introduce objetivos y familiariza al jugador con el entorno virtual y las tareas, que incluyen interactuar con clientes, tomar pedidos, recordar y seleccionar opciones del menú y gestionar la entrega de alimentos a las mesas, con la idea de simular situaciones que requieren de habilidades ejecutivas como planificación, toma de decisiones y adaptación a cambios.

Figura 4. Interfaz de Neuronat. En esta imagen se aprecia el personaje jugable (el mesero) además de cuatro mesas con clientes a los cuales se les debe tomar pedido, llevar comida y pedirles la cuenta. Fuente: Neuronat.

Las situaciones están imbuidas de elementos socioemocionales, como los roles sociales de los personajes o enfrentar dilemas que involucran emociones y valores personales, tal como muestra la tabla 3.

Dominios Cognitivos	Tarea asociada en Neuronat	Neuronat
Planificación	Evaluar el tiempo que toma al jugador atender una mesa y el número de tareas completadas correctamente. Esto incluye la evaluación de la trayectoria seguida por el sujeto para ir de la cocina a una mesa	
Memoria de trabajo	Cantidad de pedidos completos realizados correctamente.	
Multitarea en contexto social	Cantidad de acciones realizadas de forma simultánea. Tiempo en realizar las acciones.	
Flexibilidad Cognitiva	Tiempo de adaptación a nuevas tareas y manejo del juego.	

Control Contextual	Respuestas frente a distintas personalidades del juego.	
Inhibición de Respuestas	Manejo de situaciones a través de dilemas.	
Memoria prospectiva	Tiempo de respuesta a eventos, número de eventos atendidos correctamente.	
Toma de Decisiones	Todas las situaciones en las que la persona evaluada debe elegir entre dos o más acciones, cada una con diferentes consecuencias	

Tabla 3. Dominios Cognitivos y cómo estos se miden en Neuronat. Fuente: Neuronat.

Neuronat recopila una cantidad significativa de datos sobre las acciones y decisiones del jugador. Esto incluye la hora a la que se realizaron, tipo y cualquier resultado asociado, pudiendo utilizarse para construir una línea de tiempo de las acciones del jugador, permitiendo un análisis detallado de su comportamiento a lo largo del tiempo.

El objetivo del juego es proporcionar un método de evaluación ecológicamente válido, para ello, el juego ha sido diseñado con un aporte multidisciplinar de profesionales para asegurarse de que las tareas del juego son relevantes y útiles para la evaluación de la disfunción ejecutiva.

En su estado actual, Neuronat ha demostrado una usabilidad y accesibilidad robustas, para atribuir los resultados de desempeño exclusivamente a las funciones ejecutivas y no a barreras técnicas. La interfaz y mecánicas del juego han sido validadas internamente por el equipo de investigación, por ejemplo, un estudio del grupo FONDEF (Olivares-Vargas et al., 2024), situaron los puntajes de la *System Usability Scale* (SUS) en rangos *Good–Excellent* en controles (87/100) y *OK–Good* en participantes con disfunción ejecutiva (63–70/100). Estos antecedentes respaldan que las métricas analizadas, como tiempos de respuesta, errores o decisiones tomadas, reflejen predominantemente las habilidades ejecutivas de los usuarios, y no dificultades derivadas del diseño o funcionamiento del software.

1.2 Problema

La evaluación de la disfunción ejecutiva (DE) enfrenta desafíos significativos. En el estudio de Martínez-Pernía et al. (2023), se señalan tres retos principales en la evaluación del funcionamiento en el mundo real:

Contenido de la evaluación: Se carece de instrumentos adecuados para investigar la complejidad del funcionamiento en el mundo real (Schmitter-Edgecombe y Arrotta, 2022), porque la FE suele evaluarse como un ámbito cognitivo aislado, desconectado del contexto social e intersubjetivo. Esto resulta en la incapacidad de detectar disfunciones ejecutivas que no se manifiestan en entornos controlados, pero sí afectan la vida cotidiana.

Estructura de la tarea: La mayoría de las pruebas de FE evalúan tareas con criterios claramente especificados. Esto contrasta con las situaciones del mundo real, que suelen carecer de criterios bien definidos para evaluar si se ha alcanzado el objetivo. Es decir, un paciente puede tener un rendimiento bueno en la entrevista y los tests que se le realizan, pero en otros contextos cotidianos que quedan fuera de la consulta médica, puede tener un pobre desempeño. Por lo tanto, se necesita desarrollar métodos de evaluación con validez ecológica que reflejen mejor las habilidades ejecutivas en situaciones de la vida real.

Medición del rendimiento: Las mediciones de los comportamientos se basan en una recogida de datos y un análisis estadístico simplista, lo cual limita la capacidad de los evaluadores para capturar la complejidad de las DE en contextos reales.

Los *Serious Games*, con su gran potencial para la evaluación de la disfunción ejecutiva, también enfrentan desafíos propios. Anderson-Hanley et al. (2011) resaltan que, aunque los juegos pueden ofrecer una evaluación detallada de habilidades ejecutivas, es esencial que estas evaluaciones sean realmente indicativas de las habilidades y que proporcionen resultados consistentes a lo largo del tiempo. La pregunta que surge con herramientas como Neuronat es si los datos recogidos reflejan efectivamente la disfunción ejecutiva en situaciones de la vida real. Cabe mencionar

que la presente tesis no analizará aspectos de aceptabilidad y usabilidad de los *Serious Games*.

En contraste, la realidad virtual (RV) enfrenta dificultades en su implementación para la detección del comportamiento en contextos reales. Según Martínez-Pernía et al. (2023), el uso de sistemas como gafas de RV puede causar ansiedad y cinetosis (mareos y náuseas), lo que afecta tanto la participación como la calidad de los datos obtenidos. Además, las interacciones sociales en la RV suelen ser poco realistas debido al "efecto del valle inquietante", donde los avatares, al parecer casi humanos, resultan incómodos para los usuarios. También existen barreras tecnológicas como el alto costo, el mantenimiento complejo y la falta de estándares aceptados, lo que limita su adopción en entornos clínicos.

Por lo tanto, mientras que los SG destacan por la riqueza y precisión de los datos obtenidos en tiempo real, la RV enfrenta desafíos significativos que complican su aplicación efectiva, especialmente en contextos clínicos donde la validez ecológica y la comodidad del usuario son esenciales.

Dado los antecedentes existentes, se considera viable desarrollar un modelo computacional de *Machine Learning* que apoye en el diagnóstico de Disfunciones Ejecutivas (DE). Existen estudios que han demostrado resultados positivos en problemas similares utilizando técnicas de ML. Sin embargo, un desafío particular que se aborda en esta tesis es la capacidad del algoritmo para identificar casos en los que los pacientes tienen un buen rendimiento en pruebas estándar de lápiz y papel, pero muestran DE en la vida cotidiana. En esta tesis se busca aplicar técnicas de ML para crear un clasificador de DE cuya validez sea comparada al juicio clínico, proporcionando una herramienta complementaria a las evaluaciones tradicionales y que tenga la capacidad de detectar DE con mayor precisión en entornos reales

2. Hipótesis

El estudio propone la hipótesis: "Un algoritmo de clasificación, basado en los datos de uso del *Serious Game* Neuronat, tiene una validez comparada al juicio clínico experto para detectar la disfunción ejecutiva". Para hacer dicha comparación se utilizarán métricas como la curva ROC y el área bajo la curva (AUC), junto con sensibilidad y especificidad.

Lo anterior fundamentado en que los SG, como Neuronat, simulan demandas cognitivas diarias, permitiendo identificar patrones de comportamiento no evidentes en evaluaciones convencionales.

3. Objetivo General

Desarrollar y validar un algoritmo de clasificación basado en los datos de uso del SG Neuronat, con el propósito de identificar disfunciones ejecutivas.

Este algoritmo aspira a ofrecer una validez comparada con el juicio clínico experto, cuya decisión está basada en el consenso del equipo clínico que toma en cuenta el rendimiento de pruebas aplicadas, entrevistas con el paciente y conocidos.

3.1 Objetivos Específicos

1. Definir la estructura con que los datos de uso del juego serán almacenados.
2. Analizar los datos de uso de Neuronat para identificar patrones que pueden indicar la disfunción ejecutiva.
3. Desarrollar algoritmos de clasificación que utilicen los datos de uso de Neuronat para predecir la disfunción ejecutiva.
4. Comparar la validez de los algoritmos desarrollados con la del juicio clínico experto.

4. Materiales y Métodos

4.1 Participantes

Este estudio se llevó a cabo con apoyo financiero FONDEF IDeA I+D 2022 #ID22I10251. El equipo FONDEF comenzó a reclutar participantes en agosto de 2023, con meta de un tamaño de muestra de 86 participantes, que se ha estimado utilizando la fórmula recomendada por van Smeden et. al. (2020)., Se realizó una selección de variables, que se ha analizado que agrupará al menos tiempos en las mesas y errores, donde se identificó por el grupo FONDEF cinco variables compuestas en los datos preliminares

$$n = exp\left(\frac{-0.508 + 0.259 \cdot \log(\Phi) + 0.504 \cdot \log(P) - \log(\text{MAPE})}{0.544}\right)$$

Donde:

- n es el número de muestras,
- Φ es la proporción del resultado de interés (en este caso, 50% por conveniencia),
- P es el número de variables del predictor, en este caso 5,
- MAPE es el *Mean Absolute Prediction Error* (Error de Predicción Absoluto Medio), donde se apunta a lograr un 10%.

Esta muestra de participantes fue equitativamente dividida, entre participantes controles y pacientes diagnosticados por consenso clínico experto del equipo FONDEF con disfunción ejecutiva, todos evaluados con Neuronat. Los lugares de reclutamiento fueron la Mutual de Seguridad y el Hospital del Salvador. Se eligió una muestra de personas hospitalizadas por covid-19 y con TEC para representar al grupo con DE. Los participantes de ambos grupos tienen entre 18 y 65 años, de ambos sexos, con al menos 6 años de educación, capacidad de dar su consentimiento informado y realizar tareas cognitivas durante al menos 20 minutos. Se excluyó a personas con deterioro cognitivo severo, impedimentos sensoriales o motores no compensados por tecnología asistida, o enfermedad sistémica discapacitante que impidiera el diagnóstico de disfunción ejecutiva.

4.2 Protocolo de Evaluación Clínica y Categorías Diagnósticas

El manual de estudio está realizado en base al protocolo clínico que desarrollado en contexto del proyecto FONDEF #ID22I10251 por un grupo multidisciplinario de investigadores clínicos (neurología, psiquiatría, neuropsicología y psicología). El protocolo incluye pasos para los pacientes con TEC, secuelados covid-19 y los controles que se describe en la Tabla 4.

La evaluación clínica fue realizada por neurólogos expertos de la Universidad de Chile y Mutual y psicólogos de Mutual o contratados para el proyecto y supervisados por los investigadores del estudio. El protocolo de evaluación permitió establecer tres categorías en pacientes controles, TEC y Covid-19: pacientes sin DE, pacientes con DE leve y paciente con DE moderada a severa de acuerdo con los rendimientos en las evaluaciones realizadas. Sin embargo, para esta tesis, las categorías usadas serán dos: pacientes sin DE y pacientes con DE.

Contenido de la evaluación		Paciente	Control
Paso 1	Entrevista por un investigador clínico debidamente capacitado para verificar el cumplimiento de los criterios de inclusión/exclusión y proceder a firmar el consentimiento informado.	Si	Si
Paso 2	Evaluación por neuropsicológico con: a. Administración de pruebas de lápiz y papel de evaluación de funciones cognitivas y ejecutivas, i) test de eficiencia cognitiva global: Evaluación Cognitiva da Montreal (MOCA) y ii) test de funciones ejecutivas: FAB	Si	Si
	b. Administración de otras pruebas de lápiz y papel de evaluación de las funciones ejecutivas: fluencias Verbales (FAS), Torre de Londres (TOL) y Batería D-KEFS	Si	Si
	c. Administración del cuestionario, Inventario del Síndrome Disejecutivo Conductual (ISDC) a un informante confiable, que evalúa los cambios relativos al comportamiento previo en doce dominios afectados en una DE y adaptado a Chile por grupo experto.	Si	No
Paso 3	Evaluación por médico neurólogo experto en neurología cognitiva mediante una entrevista clínica aplicada a paciente e informante.	Si	No

Paso 4	Diagnóstico por juicio clínico experto: consenso entre médico experto y neuropsicólogo para establecer el diagnóstico considerando los resultados de las evaluaciones.	Si	No
Paso 5	Evaluación con el SG: de forma remota o presencial siguiendo las instrucciones desarrolladas en los proyectos previos. Para asegurar el ciego en el ensayo clínico observacional de prueba diagnóstica el SG será administrado por un evaluador que no participe del proceso de evaluación clínica por juicio experto.	Si	Si

Tabla 4. Protocolo de Evaluación Clínica. Fuente: FONDEF IDeA I+D 2022 #ID22I10251

Criterios comunes para toda la muestra:

Criterios de inclusión:

i) Edad entre 18 y 65 años,

ii) ambos sexos,

iii) mínimo de 6 años de escolaridad,

iv) disponibilidad de un informante confiable que pueda reportar sobre su comportamiento y funcionamiento en la vida diaria (que lo vea al menos 2 horas por semana), para aplicar cuestionario DEX Informante.

v) capacidad para consentir su participación en el estudio y firmar el consentimiento informado,

vi) capacidad para realizar tareas cognitivas por al menos 20 minutos.

Criterios de exclusión:

i) Trastorno cognitivo severo según escala de screening cognitivo telefónico (*MoCA blind*: persona que no es capaz de contestar o con puntaje bajo, que es discutido en reunión clínica y se decide excluir),

ii) trastorno sensorial (visión y audición) no compensado con ayudas técnicas o motor de la mano dominante o del lenguaje que interfieran con la evaluación,

iii) enfermedad sistémica invalidante causante de discapacidad qué no permita establecer con certeza el diagnóstico de una DE.

Criterios adicionales y exclusivos para controles

Criterios de inclusión:

i) Ausencia de trastornos neurológicos o psiquiátricos mayores;
ii) residentes de la comunidad y capaz vivir de manera independiente,
iii) rendimiento en rangos normales en escala de disfunción ejecutiva *Frontal Assessment Battery* (FAB) versión chilena

Criterios de exclusión:

i) Antecedentes de enfermedad neurológica y/o psiquiátrica mayor, abuso de alcohol o drogas, uso de psicofármacos, o trastorno depresivo o ansioso.

Criterios adicionales y exclusivos para pacientes (TEC)

Criterios de TEC para proyecto Mutual del trabajador

i) TEC que haya requerido consulta a servicio de urgencia u hospitalización con tiempo mayor de 6 meses desde el TEC
ii) Capacidad de lecto-escritura funcional

Criterios adicionales y exclusivos para pacientes (COVID)

Criterios de inclusión:

i) Antecedente clínico de infección por COVID-19 que haya requerido hospitalización, con auto/hetero reporte de al menos un signo neurológico/síntoma de disfunción ejecutiva
ii) Tiempo mayor de 3 meses desde la hospitalización por COVID-19
iii) Capacidad de lecto-escritura funcional
iv) Pacientes de OE1 deben tener disfunción ejecutiva (leve a moderada).

Criterios de exclusión:

i) Discapacidad intelectual premórbida sin acceso a educación formal.

El reclutamiento se realizará en coordinación con las instituciones médicas correspondientes y los participantes proporcionarán su consentimiento informado antes de participar en el estudio.

4.3 Normas y reglamentaciones pertinentes y aplicables al proyecto

Como esta tesis se encuentra enmarcada en un proyecto FONDEF, se adherirá a normativas éticas y legales tanto nacionales como internacionales a las cuales está suscrito este último, es decir, seguirá las directrices de la Declaración de Helsinki de la Asociación Médica Mundial y cumplirá con legislaciones chilenas relevantes, incluyendo la Ley N°20.584 sobre derechos y deberes en atención de salud y la Ley N°20.120 que regula la investigación científica en humanos y su genoma. En cuanto al almacenamiento de datos, se respetarán las disposiciones de la Ley N°19.628 de Protección de Datos Personales de Chile y se seguirán estándares internacionales como HIPAA y GDPR para el desarrollo de software y procesos. Antes de iniciar la investigación, los protocolos y consentimientos informados serán presentados para aprobación al comité de ética del SSMO.

4.4 Herramientas y Técnicas Utilizadas

En esta sección se describen las herramientas y técnicas empleadas en la investigación para el desarrollo y análisis de datos en el diagnóstico de disfunción ejecutiva.

Python

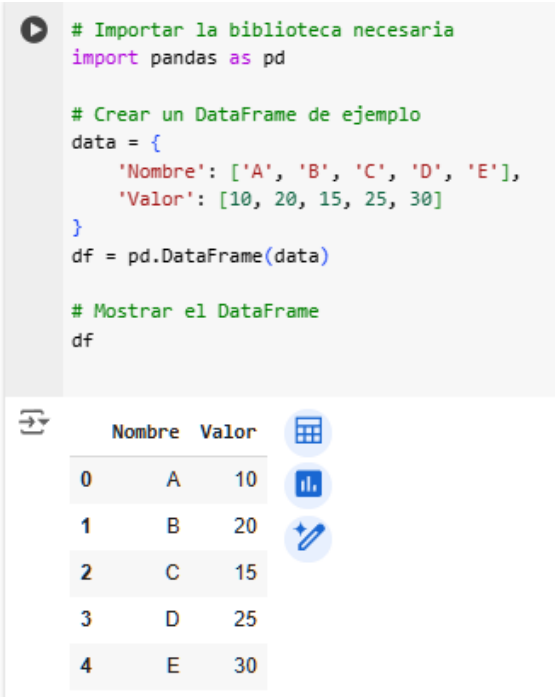
Python es un lenguaje de programación de alto nivel ampliamente utilizado en la ciencia de datos debido a su versatilidad y amplia gama de bibliotecas, con funcionalidades específicas según el problema que se esté resolviendo (Van Rossum & Drake, 2009). Python está basado en OOP (*object-oriented programming*, o programación orientada a objetos), lo que permite la creación de estructuras reutilizables y modularizadas de código, facilitando el mantenimiento y la escalabilidad de los proyectos. Su sintaxis clara, como se observa en la Figura 5 y su fuerte soporte comunitario lo hacen adecuado para la rápida prototipación y desarrollo de soluciones en el ámbito de la investigación científica.

```python
# Importación de la librería
import numpy as np

# Creación de un objeto (array)
array = np.array([1, 2, 3, 4, 5])

# Aplicación de un método (calcular la media del array)
mean_value = array.mean()

print(f"La media del array es: {mean_value}")
```

Figura 5. Importación de una librería, creación de un objeto y aplicación de método. Fuente: Elaboración propia

Jupyter Notebook

Es un entorno de desarrollo interactivo que permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo, como se observa en el ejemplo de la Figura 6. Este entorno es especialmente útil para el análisis de datos exploratorio, la visualización y la creación de informes reproducibles (Kluyver et al., 2016). La capacidad de integrar código y resultados en un solo documento facilita la colaboración y la replicabilidad de los experimentos.



```python
# Importar la biblioteca necesaria
import pandas as pd

# Crear un DataFrame de ejemplo
data = {
    'Nombre': ['A', 'B', 'C', 'D', 'E'],
    'Valor': [10, 20, 15, 25, 30]
}
df = pd.DataFrame(data)

# Mostrar el DataFrame
df
```

	Nombre	Valor
0	A	10
1	B	20
2	C	15
3	D	25
4	E	30

Figura 6. Ejemplo de la programación en Jupyter Notebook. Fuente: Elaboración propia

JSON (*JavaScript Object Notation*)

JSON es un formato de intercambio de datos ligero y de fácil lectura tanto para humanos como para máquinas. Es ampliamente utilizado para almacenar y transportar datos debido a su estructura simple y su capacidad para representar datos complejos como arrays y objetos, en la Figura 7, se muestra un ejemplo generado por Neuronat. En esta tesis, JSON se utiliza para almacenar los datos generados por los juegos serios, permitiendo una manipulación eficiente y un fácil acceso a los datos almacenados (Crockford, 2006).

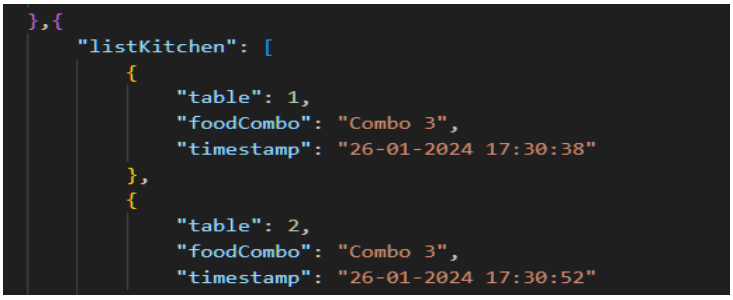


Figura 7. Ejemplo JSON generado por Neuronat, en este caso se puede ver qué pedido realizó a las mesas 1 y 2. Fuente: Elaboración propia

Machine Learning y Modelos de Clasificación

El aprendizaje automático (ML) es un campo de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para permitir a las computadoras realizar tareas específicas sin instrucciones explícitas. Existen distintos tipos de modelos de *Machine Learning*, dependiendo del tipo de entrenamiento utilizado:

- **Aprendizaje Supervisado**: Este enfoque utiliza datos etiquetados para entrenar el modelo, donde cada entrada tiene una salida correspondiente. Es comúnmente empleado para problemas de clasificación y regresión. Ejemplos de algoritmos supervisados incluyen *Support Vector Machines* (SVM), *Random Forest* y *K-Nearest Neighbors* (KNN)

- **Aprendizaje No Supervisado**: A diferencia del aprendizaje supervisado, estos algoritmos trabajan con datos no etiquetados, buscando patrones o estructuras ocultas en los datos. Ejemplos incluyen algoritmos de *clustering* como *k-means* y de reducción de dimensionalidad como PCA (Análisis de Componentes Principales), un ejemplo de esto se observa en la Figura 8:

Figura 8. Ejemplo de Clustering en Aprendizaje no supervisado, la agrupación detecta patrones ocultos en los datos, Fuente: Matlab, How to Works Machine Learning

- **Aprendizaje Semi Supervisado**: Combina datos etiquetados y no etiquetados para entrenar el modelo, aprovechando grandes cantidades de datos no etiquetados junto con una pequeña cantidad de datos etiquetados para mejorar el rendimiento del modelo.

- **Aprendizaje por Refuerzo**: Este tipo de ML entrena modelos para tomar decisiones basadas en la retroalimentación de sus acciones en un entorno dinámico. Se utiliza en áreas como la robótica y los juegos, donde los algoritmos reciben recompensas o castigos según sus acciones.

El proceso general de creación y entrenamiento de un modelo de ML implica dividir la base de datos en subconjuntos de entrenamiento y prueba. El subconjunto de entrenamiento se utiliza para enseñar al modelo a detectar patrones, que luego se aplican para clasificar o predecir nuevos datos en el subconjunto de prueba. La eficacia del modelo se evalúa comparando sus predicciones con los valores reales, minimizando el error de predicción para mejorar el rendimiento del modelo.

En esta tesis, se aplican modelos de clasificación supervisados para analizar y clasificar los datos de los pacientes.

Algoritmos de Clasificación

En el ámbito del ML, existen múltiples algoritmos de clasificación, cada uno con sus respectivas ventajas y desventajas (Bishop, 2006). A continuación, se presentan algunos de los algoritmos más utilizados en esta área:

- **Árboles de Decisión**: Este algoritmo funciona determinando una serie de reglas para las variables del vector de entrada y, según se cumplan dichas reglas, clasifica la observación en una clase específica. Los árboles de decisión son intuitivos y fáciles de interpretar, y pueden ser construidos de manera dinámica utilizando diversos criterios de división (Quinlan, 1986).

- **Regresión Logística (*Logistic Regression*)**: Es uno de los modelos lineales más utilizados en clasificación binaria. Estima la probabilidad de pertenencia a una clase aplicando la función logística (sigmoide, como en la Figura 9) sobre una combinación lineal de las variables predictoras. Aunque es un modelo simple y fácil de interpretar, su capacidad predictiva puede verse limitada en problemas no lineales o con alta complejidad (Hosmer et al., 2013). Sin embargo, cuando se combina con técnicas de regularización o selección de variables, puede ofrecer resultados competitivos.

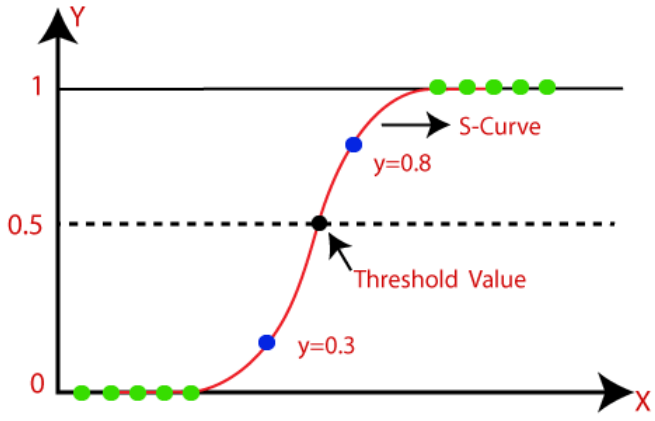


Figura 9. Representación del Modelo de Regresión Logística, Fuente: Arista (2022)

- **Random Forest**: Este algoritmo combina múltiples árboles de decisión para mejorar la precisión de la clasificación. Cada árbol se construye a partir de una muestra *Bootstrap* de los datos, y la clasificación final se determina por mayoría de votos de los árboles individuales, como se ve en la Figura 10. Random Forest es robusto contra el sobreajuste y es eficaz para manejar grandes conjuntos de datos con muchas variables (Breiman, 2001).

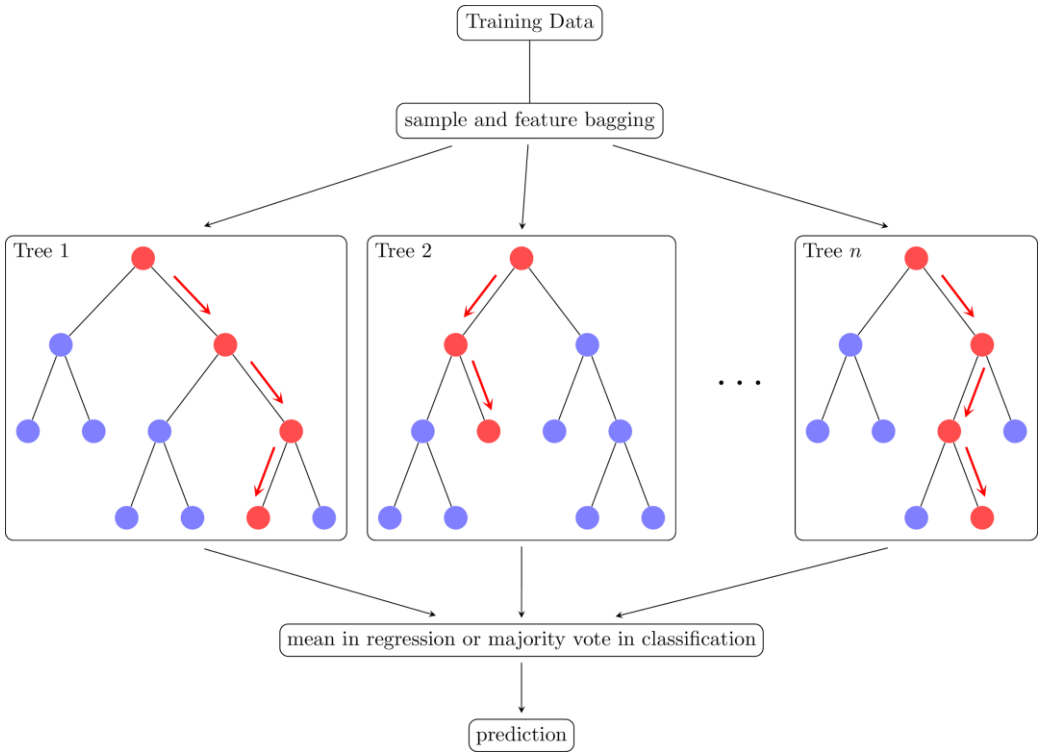


Figura 10. Diagrama del algoritmo de bosque aleatorio (RF) (Breiman 2001), los modelos de conjuntos formados por árboles de decisión binarios que predicen la moda de las predicciones de los árboles individuales en clasificación o la media en regresión.

- **Gradient Boosting (GB)**: Es una técnica de ensamblado que construye modelos aditivos de forma secuencial, entrenando cada nuevo árbol para corregir los errores del modelo anterior. Esto se realiza minimizando una función de pérdida mediante gradiente descendente, lo que permite una mejora continua del rendimiento. A pesar de ser más lento que otros métodos como Random Forest, suele ofrecer una mayor precisión en problemas complejos (Friedman, 2001).

- **XGBoost (Extreme Gradient Boosting)**: Es una implementación optimizada del algoritmo de Gradient Boosting que incluye mejoras en velocidad, regularización y manejo de valores faltantes. Está diseñado para eficiencia computacional y rendimiento, siendo ampliamente usado en competiciones de ciencia de datos y en contextos clínicos por su capacidad para manejar datasets con muchas variables y relaciones no lineales (Chen & Guestrin, 2016).

- **Support Vector Machine (SVM)**: SVM es un algoritmo que aprende a clasificar datos a través de la optimización de un hiperplano que separa las diferentes clases, esto se esquematiza en la Figura 11. Utiliza conceptos como el margen máximo del hiperplano, márgenes suaves y la función Kernel para manejar datos que no son linealmente separables (Cortes & Vapnik, 1995). SVM es eficaz para problemas de clasificación binaria y puede ser adaptado para problemas multiclasificación.

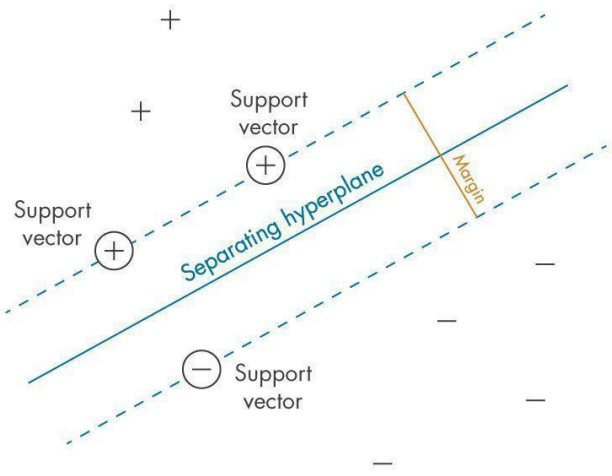


Figura 11. En la mayoría de los problemas prácticos, SVM maximiza el margen flexible permitiendo un pequeño número de clasificaciones erróneas. Fuente: Matlab, Introducción a Support Vector Machine

Métodos de Evaluación de Algoritmos de Clasificación

La evaluación de los algoritmos de clasificación se llevó a cabo utilizando técnicas como la validación cruzada, las matrices de confusión y métricas de rendimiento como precisión, exhaustividad y especificidad. Estos métodos permiten determinar la eficacia y la confiabilidad de los modelos en la predicción de disfunciones ejecutivas, proporcionando una base sólida para su implementación clínica (Fawcett, 2006).

Para medir y comparar el desempeño de los modelos, se utilizan varias métricas de evaluación. A continuación se presentará el detalle de dichas métricas para evaluar los modelos de clasificación, pero antes es importante explicar algunos conceptos. Para simplificar la explicación se considera un problema de clasificación binaria, es decir, se tienen 2 posibles clases:

- **Verdaderos Positivos (TP)**: Número de instancias en las que se predice correctamente la clase 1.
- **Falsos Positivos (FP)**: Número de instancias en las que se predice incorrectamente, retornando la clase 2 cuando el valor real era la clase 1.
- **Verdaderos Negativos (TN)**: Número de instancias en las que se predice correctamente la clase 2.
- **Falsos Negativos (FN)**: Número de instancias en las que se predice incorrectamente, retornando la clase 1 cuando el valor real era la clase 2.

Estos valores se representan en una matriz de confusión como se observa en la Figura 12, que permite simplificar la interpretación de los resultados del modelo y construir métricas de rendimiento.

		Real	
		Clase 1	Clase 2
Predicción	**Clase 1**	TP	FP
	Clase 2	FN	TN

Figura 12. Matriz de confusión de 2 clases. Fuente: Elaboración propia

Métricas de Evaluación:

Accuracy (Exactitud): Representa la proporción de clasificaciones correctas sobre el total de clasificaciones. Se calcula como:

$$\text{accuracy} = \frac{\text{Total de clasificados correctamente}}{\text{Total de registros}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (Precisión): Indica la proporción de valores positivos predichos correctamente sobre el total de valores positivos predichos. Se calcula como:

$$\text{precision} = \frac{\text{Total de clasificados correctos para la clase positiva}}{\text{Total de clasificados a la clase positiva}} = \frac{TP}{TP + FP}$$

Recall (Sensibilidad): Indica la probabilidad de que el modelo detecte casos realmente positivos. Se calcula como:

$$\text{recall} = \frac{\text{Total de clasificados correctos para la clase positiva}}{\text{Total de registros de la clase positiva}} = \frac{TP}{TP + FN}$$

F1-Score: Combina precisión y recall para proporcionar una medida del desempeño general del modelo. Se calcula como:

$$\text{F1-Score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Curva ROC y AUC: La curva ROC (*Receiver Operating Characteristic*) grafica la tasa de verdaderos positivos (TPR o recall) contra la tasa de falsos positivos (FPR o fallout). El área bajo la curva (AUC) es una medida del desempeño general del modelo, donde valores más cercanos a 1 indican mejor rendimiento, como se señala en la Figura 13.

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

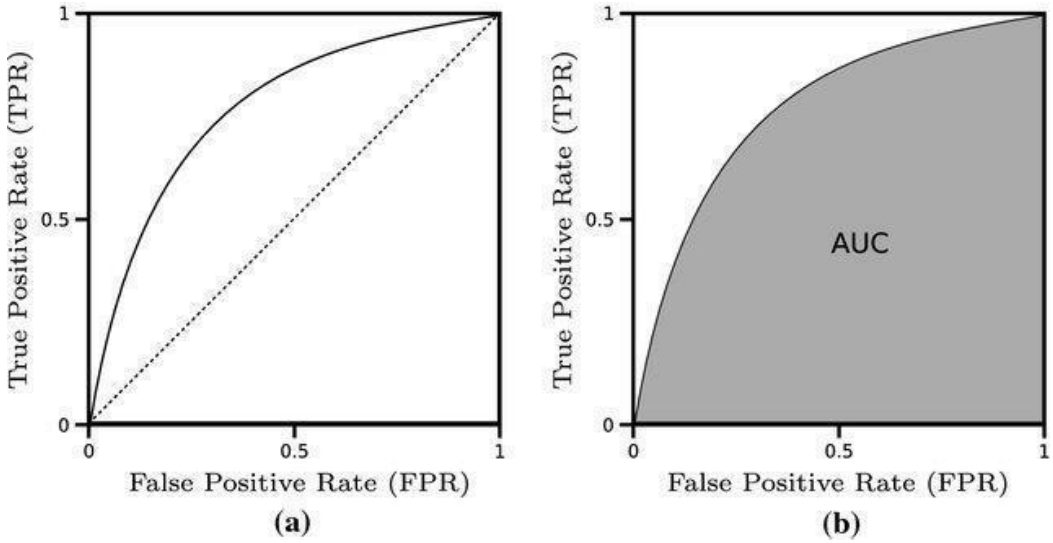


Figura 13. (a) Una curva ROC se puede obtener a partir de un clasificador de puntuación. La línea discontinua en (a) corresponde a un clasificador con un rendimiento comparable al de un clasificador completamente aleatorio. (b) El Área Bajo la Curva (AUC) de la curva correspondiente está resaltada. Fuente: Jaskowiak et. al., 2022

Estrategias de Validación

Para estimar de forma robusta estas métricas, se emplearon dos estrategias de validación:

- **Validación Cruzada k-Fold (*K-fold Cross Validation*):** El conjunto de datos se divide en k subconjuntos del mismo tamaño. El modelo se entrena k − 1 veces y se valida en el subconjunto restante, rotando hasta que todos los subconjuntos han sido utilizados como conjunto de prueba. Esta técnica permite evaluar el rendimiento del modelo en múltiples particiones, reduciendo la dependencia de una sola división.
- **Validación Cruzada Anidada (*Nested Cross-Validation*)**: Para evitar el sobreajuste durante el ajuste de hiperparámetros, se aplicó validación cruzada anidada. Este método combina una validación cruzada externa (para evaluar el rendimiento general del modelo) y una validación cruzada interna (para optimizar los hiperparámetros). De esta manera, se evita que la selección del modelo interfiera con su evaluación, proporcionando estimaciones más realistas y menos sesgadas del rendimiento.

4.5 Estructura de almacenamiento de los datos de uso del juego

El primer objetivo específico de la tesis es definir la estructura en la que se almacenarán los datos de uso del juego. Se propone una estructura JSON, que es un formato abierto y de texto ligero para el intercambio de datos, como se puede observar en la Figura 14. Esta estructura identifica componentes de interacción del jugador con el juego, incluyendo campos, la frecuencia de ciertas acciones, las posiciones en el juego, los errores cometidos y la temporalidad de estas acciones y eventos. Este formato permitió calcular y analizar métricas de juegos que se han usado en *Serious Games* creados como herramientas diagnósticas, así como algunas que serán propuestas específicamente por el equipo de investigación (El-Nasr, Drachen, & Canossa, 2013).


```
{
    "listGiveEntry": [
        {
            "food": 10,
            "character": "Hombre Solo",
            "timestamp": "02-05-2023 18:48:09"
        },
        {
            "food": 1,
            "character": "Hombre Solo",
            "timestamp": "02-05-2023 18:48:11"
        }
    ],
```


Figura 14. Estructura JSON de la entrega de un pedido en el juego.

El detalle de las métricas ocupadas se encuentra en el Anexo A. Algunos ejemplos de las métricas usadas incluyen:

- **Tiempo de juego**: Duración total que un jugador pasa en el juego.
- **Errores cometidos**: Número de errores o decisiones incorrectas tomadas por el jugador.
- **Tareas completadas**: Número de tareas o misiones completadas en el juego.
- **Interacciones**: Cantidad y tipo de interacciones con otros elementos o personajes del juego.

- **Decisiones tomadas**: Elecciones realizadas por el jugador que afectan el resultado del juego.
- **Retroalimentación**: Respuestas del juego a las acciones del jugador, como recompensas o penalizaciones.

4.6 Identificación de características para la Disfunción Ejecutiva

Se definió un conjunto de características de interés basado en la literatura existente donde ya se han aplicado SG para el diagnóstico de otras enfermedades y la experiencia de los expertos del equipo FONDEF. Estas características incluyeron, por ejemplo, el tiempo que el jugador tarda en realizar ciertas acciones, el número y tipo de errores que comete y su habilidad para realizar varias tareas a la vez. Además, los datos se representarán como un grafo y se extraerá información de este espacio de representación mediante el cálculo de diversas métricas (Rossi & Ahmed, 2015). La representación en forma de grafo implica visualizar los datos como nodos y aristas, donde los nodos pueden representar jugadores, acciones o eventos, y las aristas representan relaciones o interacciones entre ellos. Las métricas sobre los grafos que se pueden calcular a partir de esta representación incluyen centralidad, densidad, modularidad, entre otras. Estas métricas proporcionan información sobre la estructura y las relaciones dentro del grafo, lo que puede ser útil para identificar patrones o anomalías en los datos. Estos métodos y métricas se pueden calcular utilizando herramientas y bibliotecas especializadas en análisis de redes, como *NetworkX* en Python.

Luego, se realizó un análisis de grupo tomando en cuenta, media, mediana, rango intercuartil, entre otros, para cada grupo. También pruebas estadísticas (*t-Student* o *U-Mann-Whitney*, dependiendo de la normalidad de los datos) para comparar las métricas del juego entre los dos grupos. Esto permitió ver diferencias significativas en el comportamiento del juego entre las personas sin disfunción ejecutiva y con disfunción ejecutiva.

4.7 Evaluación de Usabilidad y Dificultad Percibida

Conforme al protocolo FONDEF #ID22I10251, al término de cada sesión de juego se aplicaron dos instrumentos estandarizados de auto-reporte. Su propósito fue verificar que los datos de desempeño analizados en esta tesis reflejaran primordialmente las funciones ejecutivas de los participantes y no problemas de interfaz o jugabilidad. Aunque dichos cuestionarios no constituyen un objetivo central de la presente investigación, sus resultados descriptivos (Sección 5.4 y Anexo F) ayudan a confirman la robustez y accesibilidad de Neuronat.

- **System Usability Scale (SUS):** escala de 10 ítems tipo *Likert* (1–5) que arroja un puntaje total entre 0 y 100; valores ≥ 68 se interpretan como buena usabilidad (Bangor et al., 2009). La Figura 15 muestra los rangos de interpretación. Esta medida capta la percepción global del sistema por parte del usuario.

- **Ítem de Dificultad Percibida:** se solicitó a los participantes calificar en una escala de 1 a 10 el nivel de dificultad percibida durante el juego, siendo 1 “muy fácil” y 10 “muy difícil”. Este indicador permitió estimar la carga cognitiva autoevaluada, complementando las métricas objetivas de desempeño.

Figura 15. System Usability Scale Scoring

Los puntajes de ambos instrumentos se analizaron y se compararon entre los grupos con y sin disfunción ejecutiva para descartar sesgos atribuidos al software. Los ítems completos y la fórmula de puntuación de la SUS se incluyen en el Anexo F

4.8 Desarrollo y evaluación de los Algoritmos de Aprendizaje Automático

Para desarrollar los algoritmos de aprendizaje automático, se extrajeron y procesaron los datos de *Neuronat*. Estos datos, que sirvieron como entrada para los algoritmos, consisten en características específicas y patrones de comportamiento del jugador, como tiempos de respuesta, decisiones tomadas, errores cometidos y secuencias de acciones. Estas características formaron en un conjunto de datos estructurado que alimentaron a los algoritmos de aprendizaje automático.

La salida del algoritmo es una clasificación binaria que indique si el jugador presenta o no disfunción ejecutiva. Se considerarán las clases diagnósticas basadas en la evaluación por el equipo FONDEF: "Sin disfunción ejecutiva", y "Con disfunción ejecutiva". Para lograr una clasificación precisa, se entrenará el algoritmo utilizando un conjunto de entrenamiento que contenga ejemplos etiquetados de cada una de estas clases.

Se combinaron las características identificadas y se evaluarán varios algoritmos de clasificación, incluyendo *Random Forest*, *Support Vector Machines* o Regresión Logística (LR). Estos algoritmos se evaluarán en términos de su sensibilidad, precisión y la medida F1, que es una combinación de precisión y sensibilidad (Bishop, 2006; Duda, Hart, & Stork, 2001).

Los algoritmos de aprendizaje automático se validarán utilizando una estrategia de validación cruzada (*Nested Cross Validation*). Esto permitirá evaluar qué tan bien el algoritmo es capaz de predecir la disfunción ejecutiva en nuevos datos. Finalmente, se comparará el clasificador utilizando métricas como la curva ROC y el área bajo la curva (AUC), junto con sensibilidad y especificidad, para cuantificar su capacidad de identificación.

5. Resultados

5.1 Estructura JSON definida

La estructura JSON propuesta, identificó campos, frecuencia, acciones, posiciones, errores, y temporalidad y validando dicha estructura calculando métricas de juegos conocidas y propuestas por el equipo, como se representa en la Figura 16:

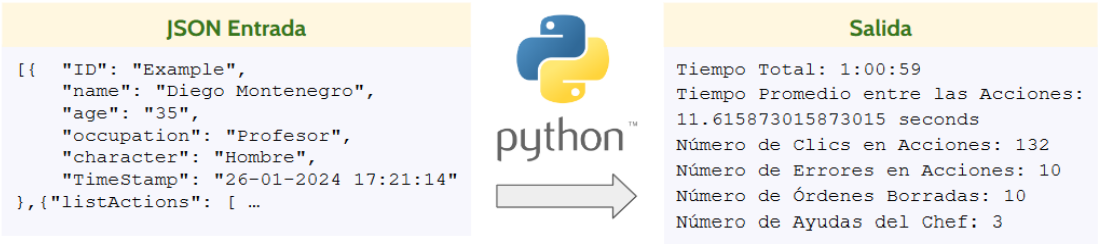


Figura 16. Esquematización del proceso de validación de la estructura JSON, usando Python como intérprete. Fuente: Elaboración Propia

Para la obtención de los datos de juego generados por *Neuronat*, se desarrolló un script en Python que permite conectarse a un servidor remoto que expone una interfaz del tipo REST API. A través del uso de solicitudes autenticadas, se accedió a los registros almacenados en formato JSON correspondientes a cada sesión de juego.

La lógica de conexión realizada en Python mediante la biblioteca *requests*, lo que permitió automatizar la descarga de los datos directamente desde el servidor hacia el entorno local de análisis. Una vez descargados, los archivos JSON fueron almacenados en carpetas clasificadas de acuerdo con la etiqueta de severidad de disfunción ejecutiva otorgada por el juicio clínico experto (Sin DE, Leve, Moderada a Severa). Esta clasificación fue fundamental para organizar los datos de entrada del clasificador y para facilitar la generación de métricas específicas por grupo. El proceso descrito, puede verse en la Figura 17:

Figura 17. Descarga de archivos JSON desde el servidor. Fuente: Elaboración Propia

El proceso de validación de la estructura JSON consistió en verificar la integridad del contenido, asegurando la presencia de campos clave como identificador de sesión, *timestamp*, lista de acciones (*listActions*) y sus atributos asociados (tipo de acción, posición, tiempo, resultado). A partir de estas estructuras se calcularon métricas propuestas por el equipo investigador, tales como número total de clics, errores por tipo, tiempo total de juego, número de ayudas solicitadas, entre otras.

Este pipeline permitió estandarizar el tratamiento de los datos de entrada, asegurando reproducibilidad y trazabilidad en las etapas posteriores de análisis y modelado.

5.2 Análisis de los datos de Uso de Neuronat

Con el objetivo de identificar patrones de uso del juego Neuronat que pudieran estar asociados a la presencia de disfunción ejecutiva (DE), se realizó un análisis exploratorio y comparativo de distintos tipos de variables extraídas a partir de los registros de interacción de los participantes.

Variables manuales de desempeño en el juego

De acuerdo con la literatura y con el equipo experto del proyecto FONDEF, se definió un conjunto de variables denominadas “manuales” reflejando dimensiones del desempeño ejecutivo de los jugadores. Estas variables incluyen:

- Tiempos de respuesta en tareas específicas.
- Errores cometidos, clasificados por tipo de acción.
- Número de tareas completadas.
- Adaptación al cambio, reflejada en decisiones tomadas frente a eventos recurrentes (como dilemas repetidos, como señala la Figura 18).

Figura 18. Ejemplo de la pregunta que se repite en dos mesas. Permitiendo comparar los resultados en situaciones contextuales distintas. Fuente: Neuronat

En la Tabla 5 se presentan algunas de estas variables, observándose diferencias estadísticamente significativas entre los grupos con y sin disfunción ejecutiva. Por ejemplo, los participantes con DE presentaron mayores tiempos totales de juego, más órdenes borradas y mayor distancia recorrida. Estos resultados sugieren patrones de interacción más ineficientes o desorganizados. El detalle de las variables Manuales se encuentra en el Anexo B.

	Severidad	Media	Test	p-valor
Tiempo total de Juego	**Sin DE**	683.30 ± 166.17	Mann-Whitney	<0.001
	Con DE	1028.7 ± 413.99		
Clics Totales	**Sin DE**	89.81 ± 21.51	Mann-Whitney	0.001
	Con DE	118.44 ± 50.85		
Errores Totales	**Sin DE**	10.81 ± 6.78	Mann-Whitney	0.027
	Con DE	12.67 ± 5.76		
Órdenes Repetidas	**Sin DE**	0.48 ± 0.96	Mann-Whitney	0.012
	Con DE	1.21 ± 1.61		
Distancia Total	**Sin DE**	232.94 ± 33.73	Mann-Whitney	0.001
	Con DE	273.92 ± 60.99		
Tiempo orden Grupo 2	**Sin DE**	17.88 ± 16.4	Mann-Whitney	0.015
	Con DE	45.0 ± 58.74		
Errores entrega Grupo 4	**Sin DE**	2.65 ± 1.94	Mann-Whitney	0.019
	Con DE	3.44 ± 1.75		

Tabla 5. Algunas variables Manuales con diferencias significativas entre grupos

Análisis de métricas basadas en grafos

Complementariamente, se utilizó un enfoque de análisis de grafos para caracterizar la secuencia de acciones realizadas durante la partida. En este enfoque, los **nodos** representan acciones individuales y las **aristas** corresponden a *transiciones* entre estas acciones, como muestra la Figura 19.

Estas métricas permitieron capturar patrones subyacentes en la ejecución de tareas, con el fin de identificar diferencias entre los grupos con y sin disfunción ejecutiva (DE).

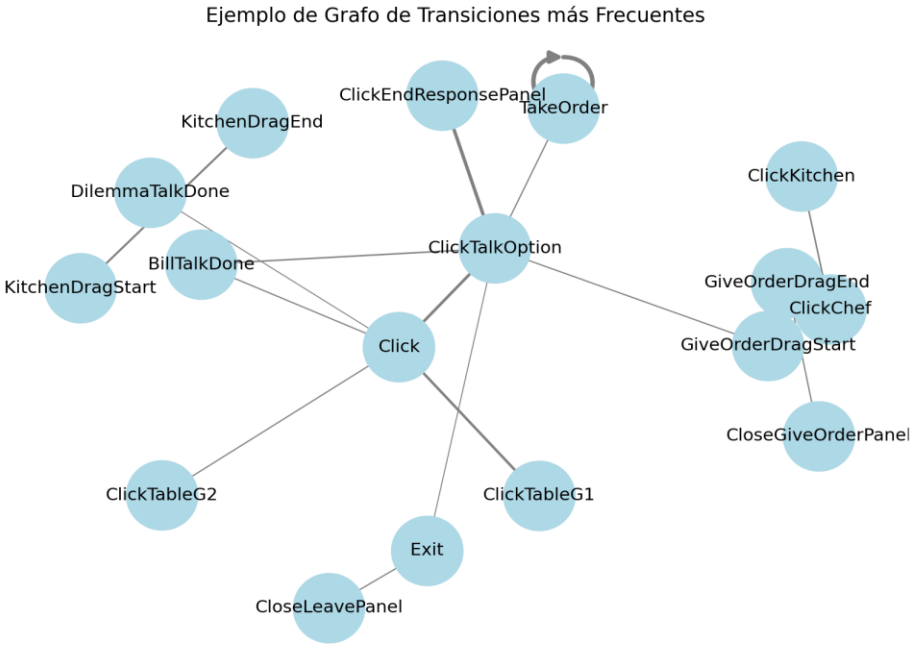


Figura 19. Ejemplo de Interacciones de acciones en un paciente. Elaboración propia

Los análisis mostraron que los participantes con DE presentan un mayor nivel de complejidad en las acciones ejecutadas, evidenciado por valores significativamente más altos en la métrica de *Action Complexity* (343.7 ± 76.34) en comparación con el grupo sin DE (311.3 ± 68.61; $p < 0.001$, prueba de Mann-Whitney). Esto sugiere que, pese a la alteración ejecutiva, los jugadores del grupo DE ejecutan secuencias de acciones más diversas durante la partida.

Asimismo, se observaron diferencias en la conectividad de los grafos de acciones. El grupo con DE presentó un *grado máximo de salida* más elevado (10.36 ± 1.559 frente a 9.559 ± 1.215; $p < 0.001$, Mann-Whitney), lo cual sugiere que existen ciertas acciones que abren paso a una mayor variedad de transiciones en los casos con disfunción.

En cuanto a la variedad general de acciones ejecutadas, también se detectó una diferencia estadísticamente significativa entre ambos grupos. La *Action Variety* fue mayor en participantes con DE (55.68 ± 2.997) que en aquellos sin DE (53.95 ± 2.875; p = 0.001, prueba de Mann-Whitney), indicando una mayor dispersión o exploración de acciones distintas por parte de quienes presentaban disfunción.

Por otra parte, el análisis de la métrica *Average Clustering,* relacionada con la repetición de ciclos locales en la ejecución de acciones, mostró una diferencia significativa entre grupos (p = 0.03, t de Student), siendo ligeramente superior en el grupo con DE (0.295 ± 0.036) frente al grupo sin DE (0.283 ± 0.038). Este patrón podría reflejar una tendencia a la repetición de ciertas secuencias por parte de quienes presentan alteraciones ejecutivas.

El resumen de las métricas más relevantes está en la tabla 6 y el detalle en el Anexo C:

	Severidad	**Media**	**Test**	**p-valor**
Complejidad de Acciones	**Sin DE**	311.3 ± 68.61	Mann-Whitney	<0.001
	Con DE	343.7 ± 76.34		
Agrupación Promedio	**Sin DE**	0.283 ± 0.038	T de Student	0.03
	Con DE	0.295 ± 0.036		
Grado Max. de Salida	**Sin DE**	9.559 ± 1.215	Mann-Whitney	<0.001
	Con DE	10.36 ± 1.559		
Variedad de Acciones	**Sin DE**	53.95 ± 2.875	Mann-Whitney	0.001
	Con DE	55.68 ± 2.997		
Grado Promedio	**Sin DE**	2.245 ± 0.128	Mann-Whitney	0.001
	Con DE	2.291 ± 0.116		

Tabla 6. Variables con diferencias significativas de las variables de Grafos de la muestra.

Análisis de desempeño neuropsicológico

Con el fin de establecer otra comparación base clínica, además del etiquetado por juicio clínico experto, se incorporaron puntajes de pruebas neuropsicológicas aplicadas a los mismos participantes. Estas variables incluyeron indicadores de funciones ejecutivas como el total del DEX (informante), desempeño en la Torre de Londres, test d2-R, MOCA y Span de Dígitos.

Los resultados muestran diferencias significativas entre los grupos en la mayoría de las pruebas, aportando una base para comparar el rendimiento del clasificador basado en datos de Neuronat. A continuación en la Tabla 7 se muestran algunas de las variables con diferencias significativas dentro de lo que son las pruebas neuropsicológicas, el detalle de las pruebas se encuentra en el Anexo D:

	Severidad	Media	Test	p-valor
Total DEX-informante (total_dex_informante)	**Sin DE**	5.41 ± 4.99	Mann-Whitney	<0.001*
	Con DE	16.86 ± 14.59		
Percentil N° de Correctas Torre de Londres (pc_tol1)	**Sin DE**	60.02 ± 28.99	Mann-Whitney	0.008*
	Con DE	42.25 ± 31.08		
Percentil Concentración d2-R (conc_pc)	**Sin DE**	52.26 ± 23.39	Mann-Whitney	<0.001*
	Con DE	32.05 ± 23.19		
Percentil Precisión d2-R (pre_pc)	**Sin DE**	53.10 ± 29.51	Mann-Whitney	0.04*
	Con DE	40.19 ± 30.95		
Span de Dígitos Directos (span_dd)	**Sin DE**	5.43 ± 0.93	Mann-Whitney	0.01*
	Con DE	4.93 ± 0.97		
Span de Dígitos Inversos (span_dd)	**Sin DE**	4.18 ± 0.97	Mann-Whitney	<0.001*
	Con DE	3.45 ± 0.93		
Total puntaje FAB (tot_fab)	**Sin DE**	16.75 ± 1.08	Mann-Whitney	<0.001*
	Con DE	14.52 ± 2.69		
Puntaje Total MoCA (puntaje_moca)	**Sin DE**	26.39 ± 2.21	Mann-Whitney	0.018*
	Con DE	24.07 ± 4.49		

Tabla 7. Algunas Variables con diferencias significativas Neuropsicológicas.

Correlación entre variables y severidad

Adicionalmente, se realizó un análisis de correlación para identificar qué variables se relacionan de forma más estrecha con la severidad de la disfunción ejecutiva. Para variables continuas se utilizó Spearman o Pearson según la normalidad de los datos, como se observa en la Figura 20, mientras que para variables categóricas y binarias se emplearon los coeficientes V de Cramer y Phi, respectivamente. Destacan con **fuerza moderada a fuerte** las siguientes correlaciones:

- **Tiempo simultáneo entre tareas** ($\rho = 0.51$; $p < 0.001$)
- **Interacciones con el grupo 2** ($\rho = 0.50$; $p < 0.001$)
- **Tiempo total de juego** ($\rho = 0.49$; $p < 0.001$)
- **Complejidad de acciones y grado de salida** también superaron valores de 0.42 y 0.47 respectivamente.

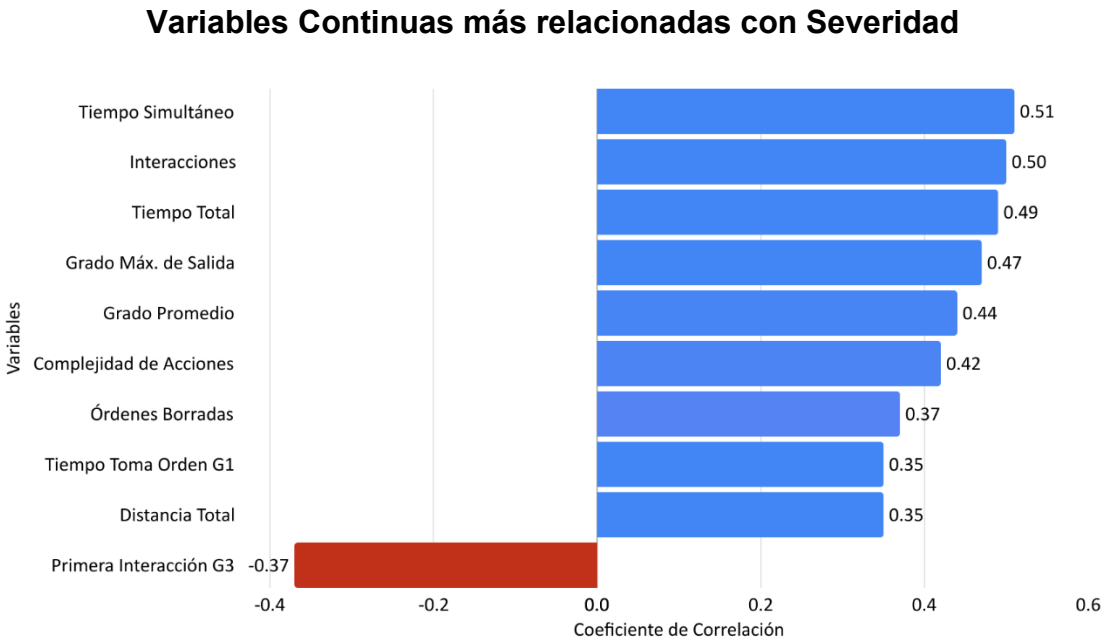


Figura 20. Top 10 variables continuas más relacionadas con Severidad.

Entre las variables categóricas, las primeras interacciones con el grupo 3 ($V = 0.384$; $p = 0.0129$) y con el grupo 4 ($V = 0.319$; $p = 0.1195$) mostraron asociaciones moderadas con la severidad, lo que sugiere un patrón temprano de comportamiento diferencial ante ciertos grupos de atención en el juego.

Estos hallazgos permitieron seleccionar variables con mayor potencial explicativo para la posterior construcción de modelos clasificadores, considerando tanto variables de desempeño general como aquellas derivadas de grafos y toma de decisiones.

Evaluación Subjetiva de la Interacción: Usabilidad y Dificultad Percibida

Con el objetivo de asegurar que las diferencias observadas en el desempeño del juego Neuronat fueran atribuibles a las capacidades ejecutivas de los participantes —y no a deficiencias en la interfaz del software—, se incorporaron dos medidas subjetivas de interacción: el cuestionario *System Usability Scale* (SUS) y un ítem de dificultad percibida. El detalle de las preguntas realizadas se encuentra en el Anexo E y F.

El cuestionario SUS, una escala estandarizada de 10 ítems que entrega un puntaje entre 0 y 100, mostró diferencias estadísticamente significativas entre grupos. Como muestra la Tabla 8, los participantes sin disfunción ejecutiva (Sin DE) reportaron un promedio de usabilidad de 81.59 ± 11.84, mientras que el grupo con disfunción ejecutiva (Con DE) reportó una media de 73.63 ± 15.93 ($U = 619$, $p = 0.008$). A pesar de esta diferencia, ambos valores se encuentran por sobre el umbral aceptable de 68 puntos, indicando que el software fue considerado usable por ambos grupos.

Respecto a la dificultad percibida, medida mediante una escala de 1 (muy fácil) a 10 (muy difícil), no se observaron diferencias significativas entre grupos. El grupo Sin DE reportó una media de 4.95 ± 2.05, mientras que el grupo Con DE reportó 5.57 ± 2.51 ($U = 769$, $p = 0.176$) , sugiriendo que ambos grupos lograron desenvolverse adecuadamente en el entorno del juego.

	Severidad	**Media**	**Test**	**p-valor**
Dificultad Percibida	**Sin DE**	4.95 ± 2.05	Mann-Whitney	0.176
	Con DE	5.57 ± 2.51		
Puntaje SUS	**Sin DE**	81.59 ± 11.84	Mann-Whitney	0.008
	Con DE	73.63 ± 15.93		

Tabla 8. Dificultad Percibida y Puntaje SUS por Severidad.

Estos hallazgos respaldan que Neuronat, es percibido como una herramienta funcional, accesible y suficientemente robusta incluso por usuarios con disfunción ejecutiva. La ausencia de diferencias significativas en la dificultad autoevaluada, junto con valores de usabilidad global por sobre el umbral aceptado, refuerzan la idea de que las diferencias en desempeño observadas entre los grupos son atribuibles a capacidades cognitivas y no a deficiencias técnicas o de diseño del software.

En la siguiente sección se presenta el desarrollo de los algoritmos de clasificación entrenados con los datos recolectados desde Neuronat, utilizando únicamente las métricas objetivas de desempeño, sin incorporar las medidas subjetivas aquí discutidas.

5.3 Desarrollo de los Algoritmos de Clasificación de Neuronat

Durante este período, se avanzó en el desarrollo y evaluación de algoritmos de clasificación para identificar disfunción ejecutiva (OE3). El cumplimiento de este objetivo incluyó los siguientes pasos:

Preprocesamiento de Datos:

Antes de entrenar los modelos se llevó a cabo un preprocesamiento de los datos extraídos de Neuronat. En primer lugar se realizó el mapeo de etiquetas de la variable objetivo, codificando la presencia o ausencia de disfunción ejecutiva (DE) como una variable binaria. A continuación se ejecutó un análisis exploratorio de datos (EDA) para estudiar la distribución de cada variable, identificar valores faltantes y detectar patrones o anomalías. Durante esta fase se eliminaron columnas redundantes y duplicados obvios. El flujo completo de este preprocesamiento de datos se representa en la Figura 21, donde se resume cada etapa clave, desde la carga de datos hasta la obtención de los conjuntos listos para modelar.

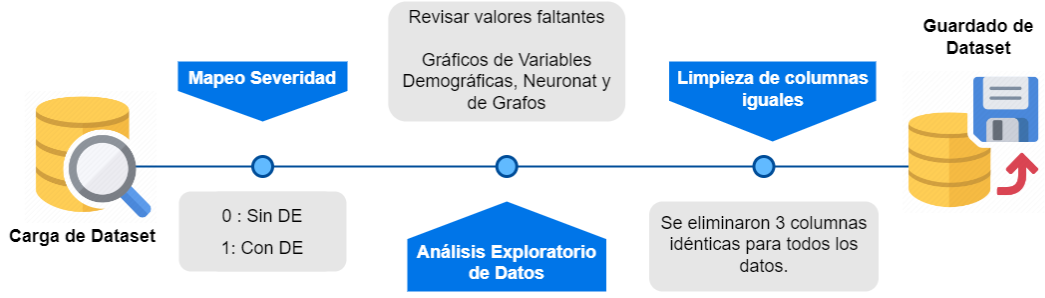


Figura 21. Proceso de Preprocesamiento de los Datos de juego de Neuronat

Selección de Variables:

Para reducir la dimensionalidad del conjunto de datos y mejorar la capacidad de generalización de los modelos, se integró un paso de filtrado de baja varianza dentro del pipeline de entrenamiento. Este filtro eliminó variables que mostraban muy poca variación entre las muestras, dado que estas no aportan información relevante al modelo.

A diferencia de otras metodologías más exhaustivas, no se aplicó en esta etapa ningún filtro basado en correlaciones ni se evaluó redundancia entre variables, ya que el enfoque se centró en conservar variables potencialmente complementarias para el aprendizaje del modelo.

Adicionalmente, en una fase previa de experimentación (evaluación no anidada), se utilizó la técnica de *SelectKBest* basada en Información Mutua (*mutual_info_classif*) como método de selección supervisada, con el fin de explorar el aporte de diferentes subconjuntos de variables al rendimiento de los modelos. Sin embargo, esta técnica no fue incorporada en la versión final del pipeline anidado.

En consecuencia, el proceso final de reducción de dimensionalidad consistió exclusivamente en la eliminación de atributos con baja varianza, ejecutada dentro del pipeline y aplicada únicamente a los datos de entrenamiento en cada pliegue de la validación cruzada anidada, para evitar fuga de información (*data leakage*).

Desarrollo de Modelos de Clasificación:

Inicialmente se entrenaron cinco modelos de clasificación supervisada dentro del conjunto de Datos Neuronat: Bosque Aleatorio (*Random Forest*), *XGBoost*, *Gradient Boosting*, *Support Vector Machine* (SVM) con kernel RBF y Regresión Logística. Para cada modelo se utilizó validación cruzada estratificada de 10 pliegues, de manera que en cada iteración se mantiene la proporción original de casos Con y Sin DE. El flujo de entrenamiento consistió en las siguientes etapas:

i) *Preprocesamiento de entrada*: escalamiento de variables continuas (por ejemplo, estandarización Z-score) y codificación *one-hot* de variables categóricas, dejando las variables binarias originales sin transformación.
ii) *Selección supervisada de características*: en cada pliegue interno, SelectKBest con información mutua seleccionó las K mejores variables predictoras.
iii) *Entrenamiento del modelo*: utilizando únicamente las características seleccionadas en ese pliegue.

Los resultados de desempeño promedio para cada modelo se evaluaron mediante métricas clave: *Recall* (sensibilidad), ROC AUC (área bajo la curva ROC) y Accuracy (precisión global). La Tabla 9 resume estas métricas comparativas obtenidas por cada algoritmo.

Modelo	Recall	ROC AUC	Accuracy
XGBoost	0.720	0.851	0.721
Random Forest	0.700	0.790	0.732
Gradient Boosting	0.700	0.807	0.729
Logistic Regression	0.685	0.828	0.744
SVM (RBF)	0.680	0.798	0.742

Tabla 9. Métricas de Evaluación Inicial de Modelos.

De forma general, los modelos basados en ensambles (Random Forest, XGBoost y Gradient Boosting) obtuvieron los mejores valores de Recall y AUC, mientras que SVM y Regresión Logística mostraron desempeño ligeramente inferior.

Dado el objetivo principal de la tesis, maximizar la sensibilidad para detectar disfunción ejecutiva, se decidió seleccionar XGBoost y Random Forest como modelos finalistas para la siguiente fase de optimización mediante *Nested Cross-Validation*. Esta elección se justifica tanto por su desempeño cuantitativo como por su robustez a múltiples métricas de evaluación.

La figura 22 muestra como la Curva ROC promedio varía entre los 10 pliegues realizados para los modelos de RF y *XGBoost* escogidos:

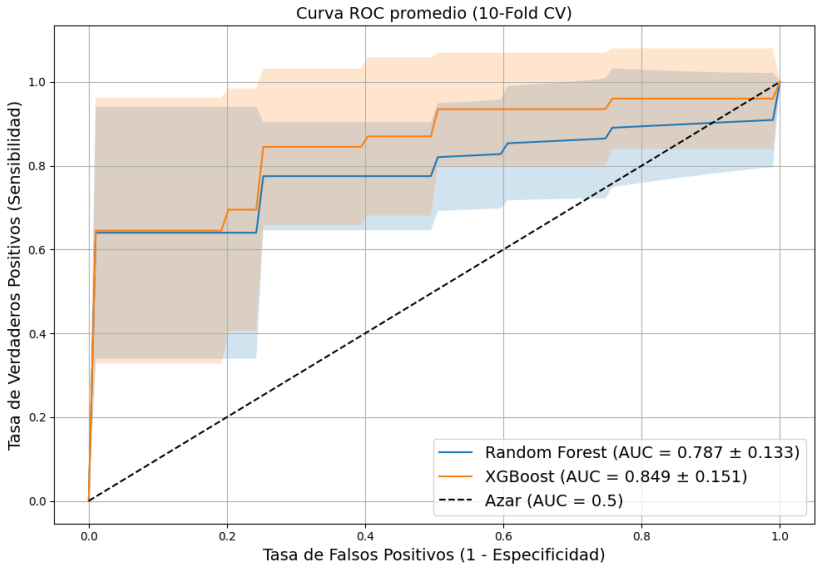


Figura 22. Curva ROC promedio de Random Forest y XGBoost.

Optimización de hiperparámetros mediante validación cruzada anidada

A diferencia de otras etapas, esta fase consideró la evaluación comparativa de modelos sobre cinco conjuntos de datos, generados a partir de la base, con combinaciones de variables, como muestra la Tabla 10:

Conjunto	Descripción	Cantidad de Variables
Neuronat	Variables las métricas del jugador, variables asociadas a los grafos y variables demográficas	59
Manual	Variables obtenidas de las métricas de Neuronat del Jugador	46
Manual +Grafos	Variables asociadas a los Grafos de Neuronat y las métricas manuales	56
Demográfico	Variables demográficas del jugador	3
Neuropsi	Variables de las evaluaciones neuropsicológicas del jugador	24

Tabla 10. Diferentes conjuntos de datos de Neuronat y Evaluaciones Neuropsicológicas

Para ajustar los hiperparámetros de los algoritmos y garantizar una evaluación imparcial, se empleó validación cruzada anidada (*nested cross-validation*) con 5 pliegues externos y 3 pliegues internos. Esta estrategia, representada en la Figura 23, implica dividir los datos en 5 bloques principales; en cada iteración uno de ellos se reserva como conjunto de prueba definitivo, mientras que los restantes 4 se usan en un proceso interno de optimización. Dentro de cada validación interna se aplicó *GridSearchCV*, priorizando la métrica de *Recall* como función objetivo. Este esquema evita sobreajuste, ya que los hiperparámetros se evalúan sobre datos no utilizados en el entrenamiento interno y no influyen en la validación externa.

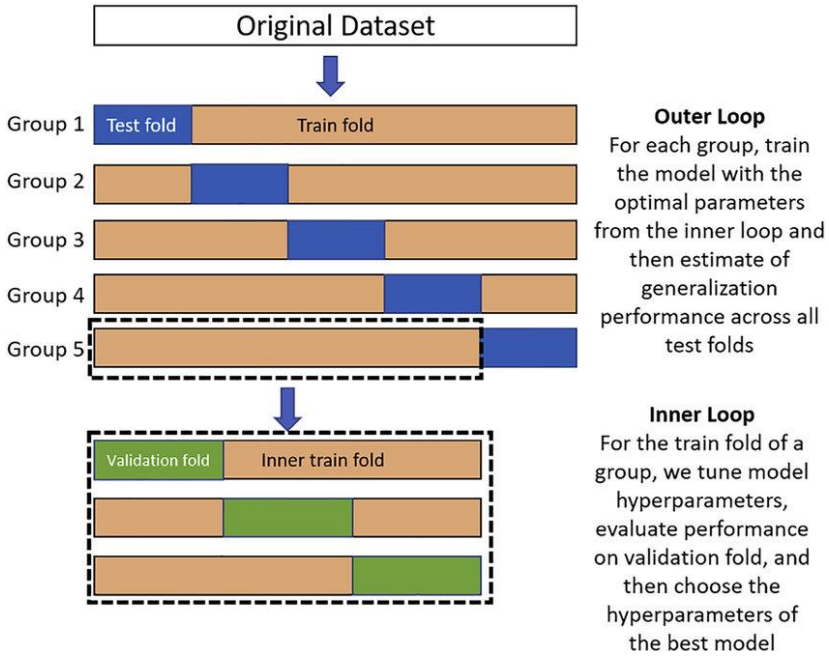


Figura 23. Representación esquemática de la validación cruzada anidada (nested cross-validation), donde el outer loop evalúa la capacidad de generalización del modelo y el inner loop se encarga de la selección de hiperparámetros. (Gupta et. al 2021)

De los modelos evaluados, *Random Forest* mantuvo una performance consistente a través de los pliegues, con poca varianza en sus métricas. Los mejores hiperparámetros encontrados incluyeron, por ejemplo, un mayor número de árboles (*n_estimators*) y profundidad moderada para *Random Forest*, así como una tasa de aprendizaje (learning_rate) reducida y mayor número de iteraciones para *XGBoost*. En general, se observó que *Random Forest* con 200 árboles y profundidad máxima en

torno a 10 obtuvo la mejor relación entre sensibilidad y especificidad. Debido a su desempeño superior y estabilidad, *Random Forest* fue seleccionado como modelo final a pesar de la competitividad de *XGBoost*.

5.4 Validez de los Algoritmos de Clasificación

Para evaluar la validez de los algoritmos de clasificación desarrollados en relación con el juicio clínico experto, se diseñó un análisis comparativo utilizando cinco conjuntos de datos. Cada uno de ellos fue utilizado para entrenar y evaluar un modelo basado en *Random Forest*, previamente optimizado mediante validación cruzada anidada, como se mencionó en el punto anterior.

Comparación de Métricas de Desempeño

En términos de exactitud (*accuracy*), los modelos entrenados con información extraída desde el serious game lograron superar claramente al modelo basado exclusivamente en variables demográficas. El modelo que alcanzó el mejor desempeño fue como se esperaba el conjunto de datos *Neuropsi* con una media de 76.9% (±11.5%), seguido por *Manual + Grafos* y *Neuronat* con valores sobre el 72%. El modelo con menor exactitud fue el Demográfico (61.0% ± 11.4%), como muestra la Figura 24:

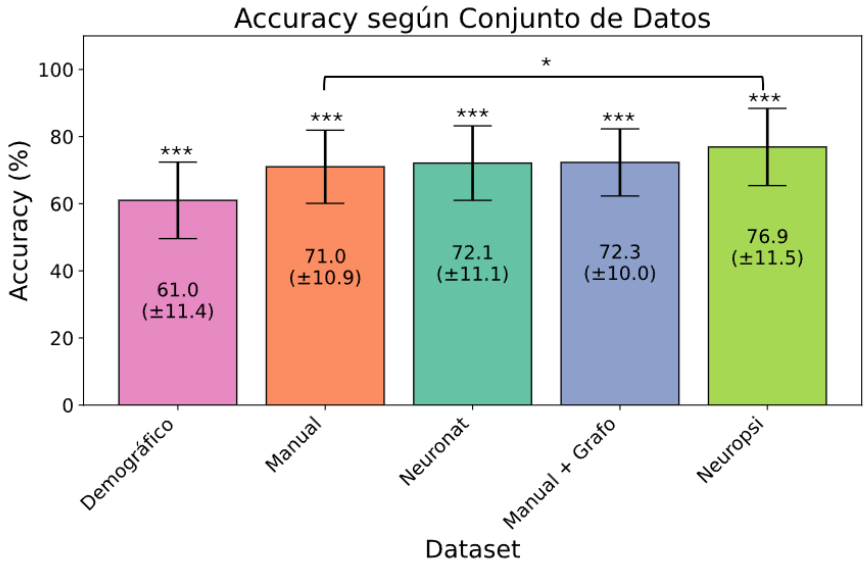


Figura 24. Comparación por Accuracy según Conjunto de Datos.

Se realizó un análisis estadístico con la prueba de Kruskal-Wallis, que reveló diferencias estadísticamente significativas entre al menos dos grupos ($p < 0.001$). Un análisis post-hoc (DSCF) identificó que el modelo demográfico difiere significativamente de todos los demás conjuntos, mientras que también se observó diferencia entre *Manual* y *Neuropsi* ($p = 0.035$).

Respecto al recall (sensibilidad), métrica prioritaria para la detección de casos positivos (con disfunción ejecutiva), los resultados fueron consistentes con los anteriores. El conjunto de datos *Neuropsi* obtuvo el valor más alto (75.3% ± 12.5%), seguido por *Manual + Grafos* (68.4%) y *Neuronat* (67.8%). El modelo *Demográfico* alcanzó nuevamente el valor más bajo (62.1%). La Figura 25 presenta esta distribución. Las diferencias fueron estadísticamente significativas (Kruskal-Wallis, $p < 0.001$), destacando diferencias específicas entre *Neuropsi* y tanto *Demográfico* ($p = 0.001$) como *Manual* ($p = 0.011$).

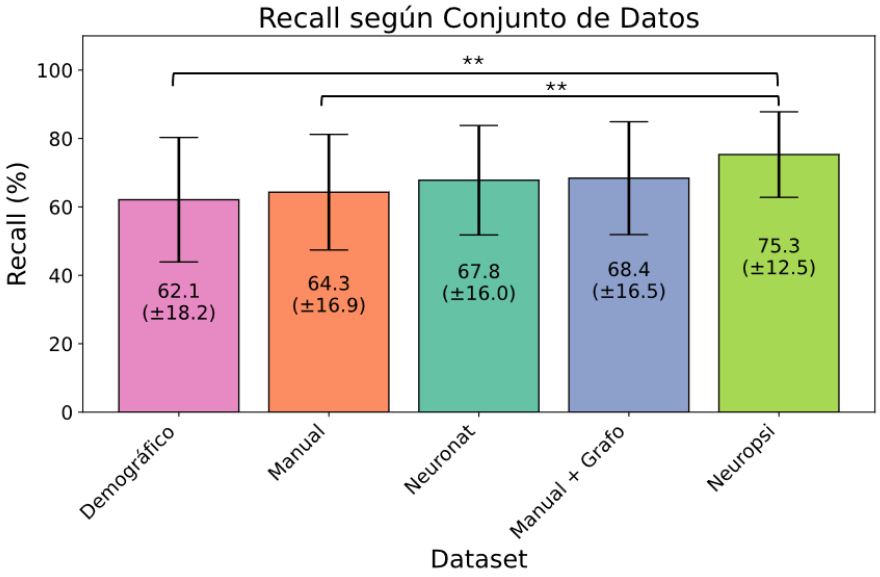


Figura 25. Comparación por Recall según Conjunto de Datos.

Además del análisis por exactitud y sensibilidad, se evaluaron métricas adicionales como el F1-score, el área bajo la curva ROC (AUC), la precisión y la especificidad, las cuales se detallan en la Tabla 11. Estas métricas permiten comprender el equilibrio entre la capacidad del modelo para detectar correctamente los casos con disfunción ejecutiva y evitar falsos positivos. En particular, el modelo basado en datos

neuropsicológicos obtuvo el mayor F1-score (0.77), seguido por Neuronat y Manual + Grafos (ambos con 0.70), lo que indica un rendimiento sólido y comparable en términos de precisión y sensibilidad combinadas.

Conjunto de Datos	F1 Score	ROC AUC	Precision	Specificity
Demográfico	0.61 ± 0.12	0.64 ± 0.13	0.61 ± 0.11	0.60 ± 0.15
Manual	0.68 ± 0.13	0.78 ± 0.12	0.75 ± 0.14	0.78 ± 0.14
Manual + Grafo	0.70 ± 0.12	0.79 ± 0.11	0.75 ± 0.12	0.76 ± 0.13
Neuronat	0.70 ± 0.12	0.78 ± 0.11	0.75 ± 0.13	0.77 ± 0.14
Neuropsicológico	0.77 ± 0.11	0.86 ± 0.11	0.80 ± 0.14	0.79 ± 0.18

Tabla 11. Rendimiento promedio según Conjunto de Datos.

Desde una perspectiva clínica, estos resultados sugieren que Neuronat logra alcanzar niveles de rendimiento similares a herramientas validadas en contexto clínico, capturando patrones relevantes de comportamiento ejecutivo a través de métricas derivadas de la interacción en el videojuego. El modelo basado exclusivamente en métricas del juego Neuronat obtuvo un *F1-score* elevado, una AUC de 0.78, y una especificidad de 0.77, lo que refuerza su capacidad para identificar correctamente tanto a pacientes con DE como a sujetos sin alteraciones, con bajo riesgo de sobrediagnóstico.

Estas evidencias apoyan la hipótesis central de este trabajo: que es posible entrenar un clasificador basado en interacción con un serious game capaz de detectar disfunción ejecutiva con validez comparable al juicio clínico experto. En este sentido, Neuronat no solo constituye una herramienta diagnóstica digital alternativa, sino que también presenta un perfil de rendimiento clínicamente relevante, con potencial para ser integrado como complemento en evaluaciones neuropsicológicas convencionales.

Importancia de Variables

Con el fin de comprender qué variables contribuían mayormente a la predicción dentro del modelo *Neuronat*, se extrajo la importancia relativa de cada característica desde el modelo *Random Forest* final entrenado. Las variables más relevantes fueron el tiempo simultáneo entre grupos activos, el tiempo total de juego, el grado promedio de conexión entre nodos del grafo, la complejidad de acciones y otras asociadas a errores o interacciones específicas. La Figura 26 muestra las diez variables con mayor importancia.

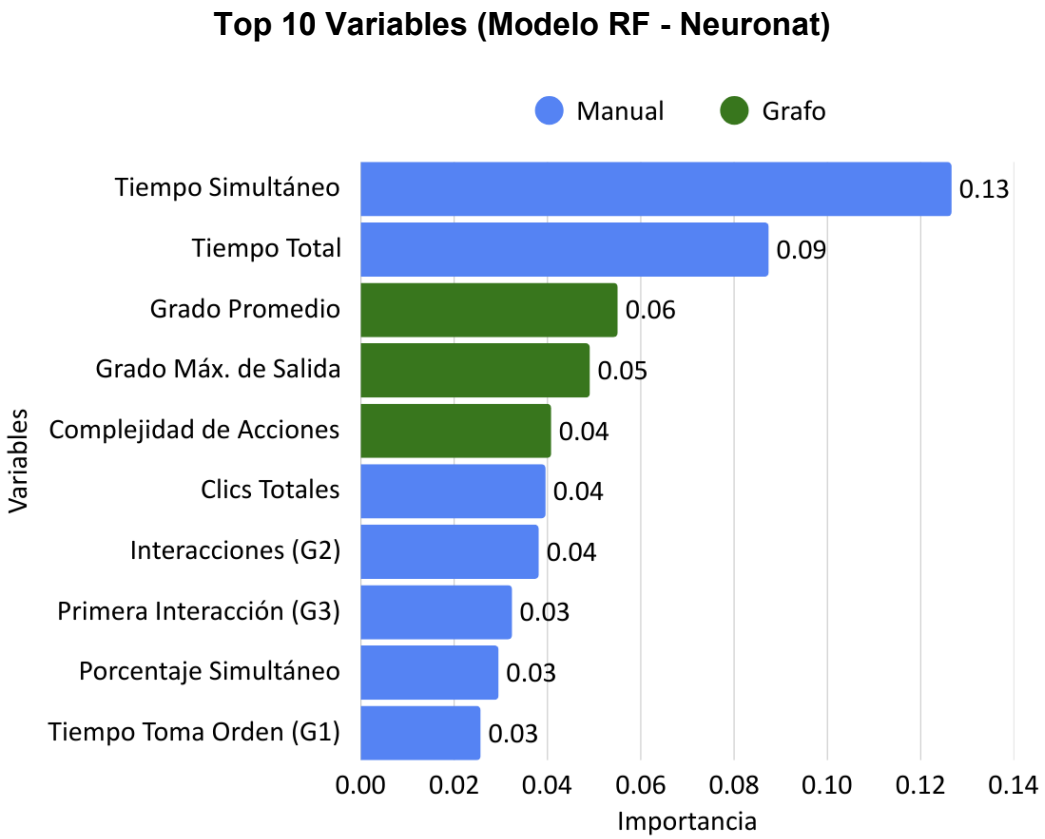


Figura 26. Variables de mayor importancia en el modelo RF para el conjunto de datos Neuronat.

Estas variables reflejan capacidades ejecutivas como planificación, eficiencia en la coordinación, y adaptabilidad frente a tareas simultáneas, todos elementos fundamentales para evaluar disfunción ejecutiva. Esto sugiere que el videojuego, sin apoyo de variables clínicas externas, puede capturar patrones relevantes en el comportamiento del jugador que reflejan alteraciones en funciones ejecutiva.

5.5 Dificultades Presentadas

Durante el desarrollo del presente trabajo se identificaron diversas dificultades que influyeron tanto en la construcción de los conjuntos de datos como en el proceso de modelado. Estas se pueden agrupar en tres categorías principales: problemas técnicos, limitaciones en la composición de la muestra y desafíos metodológicos asociados a la clasificación de casos leves.

a) Codificación de Partidas Guardadas

Una de las principales dificultades técnicas estuvo relacionada con la estructura de los datos entregados por el videojuego Neuronat. La programación del juego fue realizada por una empresa externa, la cual no implementó un sistema estandarizado ni documentado para el almacenamiento de las partidas. Esto implicó que, al momento de analizar los archivos JSON, se debiera realizar una revisión manual de las estructuras internas, reconstruyendo las secuencias de acciones y validando la presencia de campos clave como *listActions*, *timestamps*, y atributos de interacción. La ausencia de documentación detallada y la necesidad de interpretar la codificación de eventos aumentó significativamente el esfuerzo requerido para la extracción y limpieza de los datos de juego.

b) Diferencias de Variables Demográficas en los Grupos

En cuanto a la composición de la muestra, se observaron desequilibrios en las variables demográficas que podrían introducir sesgos en los resultados. Específicamente, fue difícil reclutar participantes con Disfunción Ejecutiva (DE) que presentaran escolaridad universitaria, así como también individuos Sin Disfunción Ejecutiva con escolaridad de educación básica o media. Esta distribución asimétrica dificulta la atribución exclusiva de diferencias en el rendimiento del juego a la presencia de DE, ya que podrían estar mediadas por el nivel educacional. Si bien se tomaron medidas estadísticas para mitigar esta influencia, la representación desigual sigue siendo un factor limitante en la generalización de los resultados.

c) Manejo de los Casos Leves

El manejo de los casos con Disfunción Ejecutiva Leve representó un desafío metodológico importante. En el enfoque binario (Sin DE vs. Con DE), la clase “Con DE” abarcaba tanto casos Leves como Moderados a Severos. Sin embargo, al implementar una clasificación multiclase para distinguir entre estos subgrupos (Sin DE, Leve, Mod-Sev), se evidenció un claro desbalance en el número de observaciones. Los casos Leves eran relativamente escasos, lo que afectó la capacidad del modelo para diferenciarlos con alta precisión.

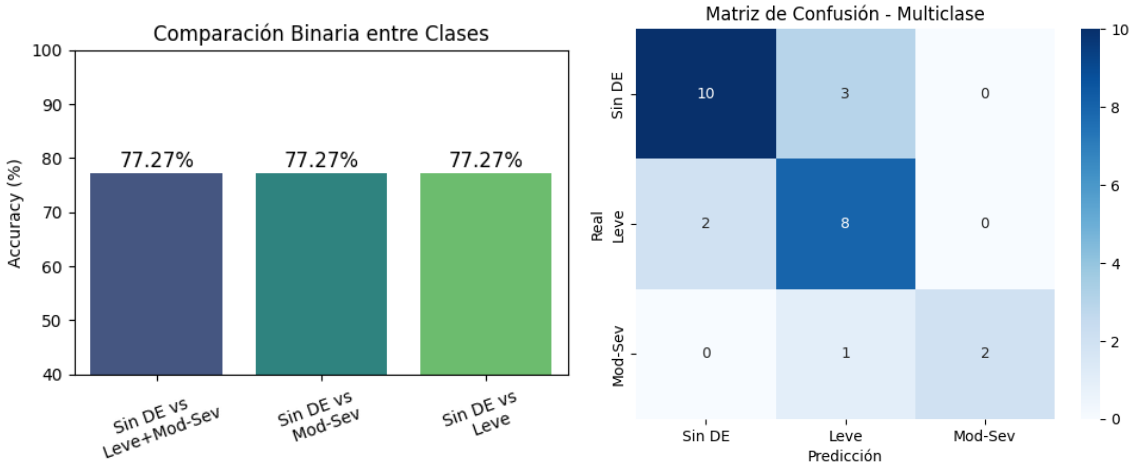


Figura 27. Rendimiento de Accuracy para un modelo multiclase en base al clasificador construido.

A pesar de ello, el modelo Random Forest entrenado sobre las métricas de Neuronat logró mantener una precisión constante del 77,27% en todas las comparaciones binarias evaluadas (Sin DE vs. Leve, Sin DE vs. Mod-Sev, Sin DE vs. Leve+Mod-Sev), como se ve en la Figura 27, lo que sugiere que las métricas extraídas del juego contienen información relevante. No obstante, la escasa cantidad de ejemplos Leves podría limitar la capacidad de generalización de los resultados hacia otras poblaciones clínicas.

6. Discusión

El análisis comparativo entre los distintos modelos y conjuntos de datos permitió extraer conclusiones relevantes sobre la capacidad predictiva de las métricas extraídas del videojuego Neuronat en relación con la disfunción ejecutiva (DE). En general, los resultados evidencian el potencial del *serious game* como herramienta de evaluación cognitiva complementaria, pero también revelan limitaciones importantes que deben considerarse en futuras aplicaciones.

6.1 Desempeño de los modelos

Los modelos construidos a partir del conjunto de datos *Neuropsi* (evaluaciones neuropsicológicas tradicionales) alcanzaron las mejores métricas en todas las dimensiones evaluadas (*Recall*, ROC AUC, *Precision, Specificity* y *F1-score*). Esto era esperable, dado que dicho conjunto representa una fuente de referencia con medidas clínicas estandarizadas.

Sin embargo, un hallazgo destacable fue el rendimiento competitivo del modelo entrenado con datos del conjunto *Neuronat*, que incluye métricas directamente extraídas del videojuego, como tiempos de ejecución, errores, ayudas solicitadas, patrones de interacción, y medidas derivadas de grafos. Este conjunto obtuvo valores sin diferencias significativas a los del *Neuropsi* en métricas como *Recall* y *Accuracy*, sugiriendo que el desempeño en el videojuego puede captar aspectos clave de la función ejecutiva.

Por otro lado, los modelos basados exclusivamente en datos demográficos o métricas manuales mostraron desempeños significativamente inferiores. En particular, el conjunto Demográfico obtuvo las peores métricas generales (*Accuracy* de 61% y ROC AUC de 0.64), indicando que, si bien edad, sexo y escolaridad son variables influyentes en los estudios de funciones ejecutivas, no son suficientes para una clasificación precisa de disfunción en este contexto.

6.2 Aportes de las variables manuales y de grafos

Un aspecto relevante fue el análisis de los conjuntos *Manual* y *Manual + Grafo*. Aunque por separado las métricas manuales mostraron un desempeño moderado, al combinarse con las variables de grafos (como complejidad de acciones o densidad de transiciones), el rendimiento del modelo mejoró en todas las métricas evaluadas. Esto sugiere que las variables de conectividad y complejidad operacional capturan aspectos complementarios del comportamiento en el juego, como la capacidad de planificación, flexibilidad y eficiencia.

El análisis de importancia de variables del modelo *Random Forest* entrenado sobre *Neuronat* mostró que los atributos más relevantes fueron: tiempo simultáneo entre grupos activos, tiempo total de juego, grado promedio del grafo de transiciones, y medidas de complejidad. Esto refuerza la hipótesis de que el desempeño estratégico y la organización de tareas en el juego reflejan componentes esenciales de la función ejecutiva.

6.3 Desafíos en la clasificación de casos leves

Uno de los mayores desafíos identificados fue el manejo de los casos leves de disfunción ejecutiva. En los experimentos multicategoría, donde se intentó distinguir entre tres clases (Sin DE, DE Leve, y DE Moderado-Severa), los modelos tendieron a confundir los casos leves con el resto. La matriz de confusión del modelo *Random Forest* mostró una precisión constante del 77.27% en clasificaciones binarias, pero un descenso en la identificación específica de la clase leve dentro del enfoque multicategoría. Esto sugiere que las métricas actuales del videojuego capturan con mayor claridad los extremos (presencia o ausencia clara de DE), pero aún presentan limitaciones para diferenciar matices leves o incipientes.

Este fenómeno podría explicarse por la naturaleza de los datos o la cantidad reducida de participantes con DE moderada-severa, que generó desequilibrio en las clases. También influye el hecho de que las funciones ejecutivas no se ven afectadas de forma homogénea y pueden variar en manifestaciones clínicas sutiles que aún no son completamente capturadas por las métricas diseñadas.

6.4 Limitaciones y proyecciones

Durante el desarrollo de este proyecto se enfrentaron diversas limitaciones que, si bien no impidieron la implementación del modelo, sí pueden haber afectado la generalización de los resultados. Una de las principales dificultades fue la codificación inconsistente de algunas partidas del videojuego. Esta situación obligó a un esfuerzo adicional para depurar y estandarizar los datos disponibles, lo que podría haber llevado a la exclusión de registros valiosos. Una solución futura contempla reprogramar el sistema de almacenamiento de eventos del juego, permitiendo una trazabilidad completa y robusta para estudios longitudinales.

También se identificaron sesgos en la distribución demográfica de los participantes: fue difícil obtener personas con DE y escolaridad universitaria, o bien personas sin DE pero con baja escolaridad. Este desequilibrio puede afectar la representatividad del modelo y debe ser considerado en futuras etapas, idealmente aumentando el tamaño y la heterogeneidad de la muestra.

Desde el punto de vista del modelo, si bien las métricas promedio fueron prometedoras, la clasificación de casos leves continúa siendo un desafío. El uso de métricas como el *Recall* permitió priorizar la detección de verdaderos positivos, pero no garantiza una óptima sensibilidad por subgrupo. Una proyección relevante es la exploración de análisis de *thresholds* personalizado, por ejemplo, ajustando el punto de corte en la probabilidad de clasificación para maximizar el F1-score o el Índice de Youden lo cual podría mejorar la detección de casos atípicos o limítrofes.

Por último, se destaca que la recolección automatizada de datos del juego, a pesar de los desafíos técnicos iniciales, demostró ser un enfoque escalable y potencialmente aplicable en contextos clínicos o de investigación. La posibilidad de estimar riesgo de disfunción ejecutiva a partir del comportamiento en un entorno lúdico computacional representa una contribución innovadora al campo de la evaluación digital.

6.5 Valor clínico y comparativo de Neuronat

Los hallazgos de esta tesis confirman lo expuesto en la Tabla 2 de Antecedentes: un serious game bien diseñado ofrece validez ecológica, granularidad de datos y alta motivación, como se explicita en la Tabla 12 pero también exhibe limitaciones que deben reconocerse (Martínez-Pernía et al., 2023). En Neuronat estas ventajas se observan incluso cuando la aplicación es supervisada por un profesional que orienta al participante sin intervenir en la resolución de la tarea.

Dimensión	**Evidencia específica de *Neuronat***	**Ventaja principal**	**Limitación identificada**
Validez ecológica	Restaurante con presión temporal y reglas sociales	Mayor transferencia a situaciones de la vida diaria	Requiere familiaridad mínima con el dispositivo y la metáfora de juego
Granularidad de datos	> 5000 eventos por sesión (tiempos, errores, grafos)	Modelos con AUC = 0.78 (similar a *Neuropsi*) usando solo variables in-game	Falta de baremos clínicos para cada métrica fina
Aplicación supervisada	Profesional entrega instrucciones y resuelve dudas puntuales	Reduce variabilidad entre examinadores y permite monitorizar incidencias	Dependencia de personal capacitado; no es 100 % auto-administrado
Motivación/ adherencia	SUS “Good–Excellent” en controles y “OK–Good” en DE	Menor fatiga que en tests papel-lápiz; baja tasa de abandono	Curva de aprendizaje digital en adultos mayores

Tabla 12. Fortalezas y limitaciones de Neuronat como herramienta de evaluación de funciones ejecutivas

Neuronat capta dominios ejecutivos complejos como planificación, multitarea, flexibilidad, con un coste tecnológico bajo y escaso tiempo de preparación. La supervisión profesional asegura estandarización y seguridad clínica, pero introduce dependencia de recursos humanos y elimina por ahora la posibilidad de aplicación totalmente remota. Además, la ausencia de normas poblacionales y la menor sensibilidad en casos leves (véase 6.3) siguen siendo retos prioritarios.

6.6 Comparación de Neuronat con estudios de referencia

Para contextualizar el rendimiento obtenido se seleccionaron cuatro estudios representativos, dos en VR, uno en 3-D no inmersivo y uno en 2-D de Tablet, cuyos dominios ejecutivos y métricas publicadas permiten una comparación directa con Neuronat. Los resultados se sintetizan en la Tabla 13:

Estudio (población)	Plataforma	Contexto / Dominio ejecutivo	Indicador clave	Observaciones
VMET – Serino et al., 2014 (34 Parkinson / 34 CTL)	**VR**	Supermercado – planificación, multitarea	Recall (0.72)	Alto costo de hardware VR; riesgo de cinetosis
V-Store – Levy et al., 2019 (≈ 200 muestra mixta)	**VR**	Tienda – planificación, memoria de trabajo	AUC (0.78)	Buena aceptabilidad; entorno totalmente inmersivo
JEF – Jansari et al., 2014 (40 adultos jóvenes)	**3-D (no inmersivo)**	Oficina virtual – inhibición, memoria de trabajo	r con WCST (0.49)	Correlación convergente; análisis de errores cualitativos
Kitchen & Cooking – Manera et al., 2015 (21 DCL/EA)	**2-D / tablet**	Cocina – planificación secuencial	r con Stroop / TMT (0.40 – 0.53)	Buen *engagement*; métrica correlacional
Neuronat – presente tesis (44 DE / 44 CTL)	**2-D / PC**	Restaurante – Multidominio	**Recall (0.68)** **AUC (0.78)**	Métricas de grafo; supervisión clínica breve; hardware convencional

Tabla 13. Rendimiento comparado de Neuronat y referencias de la literatura. Se incluyen estudios con métricas publicadas (sensibilidad, AUC o correlaciones convergentes) que evalúan dominios ejecutivos semejantes.

En los cinco entornos analizados, las variables tiempos de respuesta y errores de regla emergen de forma reiterada como los predictores más robustos de disfunción ejecutiva. Además, aunque la intensidad de la evidencia varía entre estudios, todos los autores coinciden en que la mayor validez ecológica de sus tareas virtuales permite detectar manifestaciones de déficit ejecutivo que las pruebas lápiz-y-papel de referencia captan con menor sensibilidad o sólo en estadios más avanzados, respaldando así la importancia de incorporar contextos realistas en la evaluación (Burgess et al., 2006).

Las principales diferencias entre los prototipos comparados se concentran en los requisitos tecnológicos y en el tipo de métrica ofrecida. Los entornos inmersivos, como V-Store y VMET, igualan el rendimiento global de Neuronat (AUC ≈ 0,78) pero exigen casco y controladores VR, lo que encarece su adopción clínica y puede inducir cinetosis. Herramientas 2-D previas, como JEF o Kitchen & Cooking, reportan correlaciones moderadas con pruebas tradicionales, pero carecen de modelos predictivos y métricas dinámicas avanzadas. Neuronat, en cambio, alcanza un Recall de 0,68 con hardware convencional y añade métricas de grafo, que explican más del 20 % de la importancia del modelo (Sección 6.2), aportando una dimensión analítica todavía ausente en los otros juegos.

7. Conclusiones

El presente trabajo abordó el desafío de evaluar la disfunción ejecutiva a través de un enfoque basado en *Serious Games*, proponiendo una metodología que combina el análisis detallado de métricas de interacción con técnicas avanzadas de aprendizaje automático. En particular, se desarrolló y evaluó un clasificador binario entrenado con variables extraídas del comportamiento de usuarios en el juego Neuronat, con el fin de discriminar entre participantes con y sin indicios clínicos de DE. Los resultados obtenidos confirman que este enfoque tiene potencial clínico relevante, especialmente como herramienta complementaria de evaluación.

Desde una perspectiva metodológica, se diseñó un pipeline robusto y reproducible, que incluyó filtrado de variables, selección supervisada basada en información mutua, codificación categórica y validación cruzada anidada para el ajuste y evaluación de múltiples modelos. Este enfoque no solo aseguró estimaciones menos sesgadas del rendimiento predictivo, sino que permitió una comparación justa entre diversos conjuntos de datos (sociodemográficos, manuales, gráficos, pruebas estandarizadas y del videojuego). La inclusión de métricas más allá de la exactitud como F1-score, recall y AUC fue importante para interpretar adecuadamente la utilidad del modelo, especialmente en contextos donde el costo de los falsos negativos es alto, como ocurre en la detección temprana de deterioro ejecutivo.

En relación con los resultados, los modelos entrenados con datos derivados del videojuego mostraron un rendimiento competitivo respecto a aquellos basados en pruebas estandarizadas o en variables sociodemográficas, con F1-scores promedio por sobre 0.70 en los mejores casos. Esto demuestra que las métricas conductuales derivadas de interacciones en entornos gamificados pueden capturar señales relevantes para la evaluación cognitiva. Particularmente, el conjunto Neuronat alcanzó un balance adecuado entre sensibilidad y precisión, destacándose variables relacionadas con el tiempo de respuesta, el manejo de órdenes y la eficiencia en la cocina virtual como elementos predictivos de disfunción.

Desde una perspectiva clínica, estos hallazgos sugieren que Neuronat no pretende reemplazar las herramientas neuropsicológicas tradicionales, sino complementarlas aportando una capa adicional de análisis basada en el desempeño cotidiano simulado. Su mayor ventaja radica en su validez ecológica, al representar actividades organizadas alrededor de un contexto funcional y relevante para la vida diaria, como lo es la gestión de un restaurante. Además, al ser una plataforma accesible, usable y funcional, se garantiza que las métricas obtenidas reflejan el desempeño del participante más que deficiencias técnicas o de diseño.

En comparación con las pruebas tradicionales, el uso de Serious Games permite recolectar grandes volúmenes de datos granulares en tiempo real, incluyendo rutas de acción, errores, tiempos de ejecución, repeticiones y decisiones bajo presión. Esta capacidad para capturar la dinámica de la conducta abre nuevas posibilidades para evaluaciones más sensibles a cambios sutiles, seguimiento longitudinal, y eventualmente para intervenciones personalizadas. No obstante, también existen desafíos, como la necesidad de validar externamente estas herramientas, establecer baremos normativos y comprender mejor la relación entre métricas de juego y constructos clínicos.

Este trabajo contribuye al debate actual sobre el rol de las tecnologías emergentes en neuropsicología clínica, al ofrecer evidencia empírica de que un Serious Game como Neuronat, estructurado adecuadamente y analizado con rigor metodológico, puede aportar valor real en contextos de evaluación. La metodología desarrollada, además, es adaptable y extensible a otras condiciones o grupos clínicos, lo que amplía su aplicabilidad en investigación y práctica profesional.

Finalmente, este estudio sienta las bases para futuras investigaciones que busquen integrar enfoques tradicionales y digitales en la evaluación de funciones ejecutivas, promoviendo herramientas más accesibles, atractivas y contextualizadas, sin perder rigurosidad diagnóstica. La validación con muestras clínicas más amplias, la integración con marcadores neurobiológicos o neuropsicológicos y la automatización de la interpretación de métricas constituyen líneas prometedoras para consolidar el uso de videojuegos serios en entornos clínicos reales.

8. Referencias

Adamit, T., Maeir, A., Ben Assayag, E., Bornstein, N. M., Korczyn, A. D., & Katz, N. (2015). Impact of first-ever mild stroke on participation at 3 and 6 months post-event: The TABASCO study. *Disability and Rehabilitation*, 37(8), 667-673. https://doi.org/10.3109/09638288.2014.923523

Anderson-Hanley, C., Arciero, P., & Snyder. (2011). Social facilitation in virtual reality-enhanced exercise: competitiveness moderates exercise effort of older adults. *Clinical Interventions in Aging*, 275. https://doi.org/10.2147/cia.s25337

Baum, C. M., Morrison, T., Hahn, M., Edwards, D., & Jenkins, L. (2008). Executive function performance test (EFPT): Test protocol for administration. Washington University School of Medicine.

Bishop, C. (2006). *Pattern Recognition and Machine Learning* (1a ed.). Springer.

Bohil, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nature Reviews. Neuroscience*, *12*(12), 752–762. https://doi.org/10.1038/nrn3122

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. M., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society : JINS*, *12*(2), 194–209. https://doi.org/10.1017/S1355617706060310

Chan, R., Shum, D., Toulopoulou, T., & Chen, E. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *23*(2), 201–216. https://doi.org/10.1016/j.acn.2007.08.010

Chaytor, N., Schmitteredgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical*

Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, *21*(3), 217–227. https://doi.org/10.1016/j.acn.2005.12.002

Chen, S., & Michael, D. (2005). *Serious games: Games that educate, train, and inform*. Course Technology

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chicchi Giglioli, I. A., de Juan Ripoll, C., Parra, E., & Alcañiz Raya, M. (2018). EXPANSE: A novel narrative serious game for the behavioral assessment of cognitive abilities. *PloS One*, *13*(11), e0206925. https://doi.org/10.1371/journal.pone.0206925

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLOS Biology*, 10(3), e1001293. https://doi.org/10.1371/journal.pbio.1001293

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Crockford, D. (2006). The application/json Media Type for JavaScript Object Notation (JSON). *Internet Engineering Task Force* (IETF).

Cummings, J. L. (1993). Frontal-subcortical circuits and human behavior. *Archives of Neurology*, *50*(8), 873–880. https://doi.org/10.1001/archneur.1993.00540080076020

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2a ed.). John Wiley & Sons.

El-Nasr, M. S., Drachen, A., & Canossa, A. (Eds.). (2013). *Game analytics: Maximizing the value of player data* (2013a ed.). Springer.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Fleming, T. M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S. N., Cheek, C., Lucassen, M., Lau, H. M., Pollmuller, B., & Hetrick, S. (2017). Serious games and gamification for mental health: Current status and promising directions. *Frontiers in psychiatry*, *7*. https://doi.org/10.3389/fpsyt.2016.00215

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Goldstein, S., Naglieri, J. A., Princiotta, D., & Otero, T. M. (2014). Introduction: A history of executive functioning as a theoretical and clinical construct. In S. Goldstein & J. A. Naglieri (Eds.), *Handbook of executive functioning* (pp. 3–12). Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-8106-5_1

Helms, J., Kremer, S., Merdji, H., Clere-Jehl, R., Schenck, M., Kummerlen, C., Collange, O., Boulay, C., Fafi-Kremer, S., Ohana, M., Anheim, M., & Meziani, F. (2020). *Neurologic features in severe SARS-CoV-2 infection*. New England Journal of Medicine, 382(23), 2268–2270. https://doi.org/10.1056/NEJMc2008597

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Jansari, A. S., Devlin, A., Agnew, R., Akesson, K., Murphy, L., Leadbetter, D., & Daggar, R. (2014). *Ecological assessment of executive functions: A new virtual reality task and its correlation with standard neuropsychological measures*. The Journal of Neuroscience, Psychology, and Economics, 7(2), 68–77. https://doi.org/10.1037/npe0000013

Jovanovski, D., Zakzanis, K., Ruttan, L., Campbell, Z., Erb, S., & Nussbaum, D. (2012). Ecologically valid assessment of executive dysfunction using a novel virtual reality task in patients with acquired brain injury. *Applied*

Neuropsychology: Adult,19(3), 207–220. https://doi.org/10.1080/09084282.2011.643956

Karapapas, C., & Goumopoulos, C. (2021). Mild cognitive impairment detection using machine learning models trained on data collected from serious games. *Applied Sciences (Basel, Switzerland)*, *11*(17), 8184. https://doi.org/10.3390/app11178184

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT Press.

Kinnunen, K. M., Greenwood, R., Powell, J. H., Leech, R., Hawkins, P. C., Bonnelle, V., Patel, M. C., Counsell, S. J., & Sharp, D. J. (2011). White matter damage and cognitive impairment after traumatic brain injury. *Brain: A Journal of Neurology*, *134*(2), 449–463. https://doi.org/10.1093/brain/awq347

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S. & Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87-90). IOS Press.

Kononenko, I. (2001). Machine Learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1), 89-109

Kourtesis, P., & MacPherson, S. E. (2023). An ecologically valid examination of event-based and time-based prospective memory using immersive virtual reality: The influence of attention, memory, and executive function processes on real-world prospective memory. *Neuropsychological Rehabilitation*, *33*(2), 255–280. https://doi.org/10.1080/09602011.2021.2008983

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology*, 21(1), 59–80. https://doi.org/10.1348/026151003321164627

Levy, C. E., Miller, D. M., Akande, C. A., Lok, B., Marsiske, M., & Halan, S. (2019). V-mart, a virtual reality grocery store: A focus group study of a promising intervention for mild traumatic brain injury and posttraumatic stress disorder. *American Journal of Physical Medicine & Rehabilitation*, *98*(3), 191–198. https://doi.org/10.1097/phm.0000000000001041

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment* (5a ed.). Oxford University Press.

Martínez-Pernía, D., Olavarría, L., Fernández-Manjón, B., Cabello, V., Henríquez, F., Robert, P., Alvarado, L., Barría, S., Antivilo, A., Velasquez, J., Cerda, M., Farías, G., Torralva, T., Ibáñez, A., Parra, M. A., Gilbert, S., & Slachevsky, A. (2023). The limitations and challenges in the assessment of executive dysfunction associated with real-world functioning: The opportunity of serious games. *Applied Neuropsychology: Adult*, 1–17. https://doi.org/10.1080/23279095.2023.2174438

McDonald, S., Flanagan, S., Rollins, J., & Kinch, J. (2003). TASIT: A new clinical tool for assessing social perception after traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, *18*(3), 219–238. https://doi.org/10.1097/00001199-200305000-00001

McGinnis, R. S., McGinnis, E. W., Hruschak, J., Lopez-Duran, N. L., Fitzgerald, K., Rosenblum, K. L., & Muzik, M. (2019). Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning. *IEEE Journal of Biomedical and Health Informatics, 23*(6), 2357-2364.

Parsons, T. D., Carlew, A. R., Magtoto, J., & Stonecipher, K. (2017). The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical neuropsychology. *Neuropsychological Rehabilitation*, *27*(5), 777–807. https://doi.org/10.1080/09602011.2015.1109524

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *37*(1), 51–87. https://doi.org/10.1111/j.1469-7610.1996.tb01380.x

Rand, D., Rukan, S. B., Weiss, P. L., & Katz, N. (2009). *Validation of the Virtual MET as an assessment tool for executive functions*. Neuropsychological Rehabilitation, 19(4), 583–602. https://doi.org/10.1080/09602010802469074

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., Gaggioli, A., Botella, C., & Alcañiz, M. (2007). Affective interactions using virtual reality: The link between presence and emotions. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *10*(1), 45–56. https://doi.org/10.1089/cpb.2006.9993

Romero-Ayuso, D., Castillero-Perea, Á., González, P., Navarro, E., Molina-Massó, J. P., Funes, M. J., … Triviño-Juárez, J. M. (2019). Assessment of cognitive instrumental activities of daily living: a systematic review. *Disability and Rehabilitation*, *43*(10), 1342–1358. https://doi.org/10.1080/09638288.2019.1665720

Rossi, R. A., & Ahmed, N. K. (2016). An interactive data repository with visual analytics. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, *17*(2), 37–41. https://doi.org/10.1145/2897350.2897355

Schmitter-Edgecombe, M., & Arrotta, K. (2022). Naturalistic assessment: Everyday environments and emerging technologies. In T. D. Marcotte, M. Schmitter-Edgecombe, & I. Grant (Eds.), *Neuropsychology of everyday functioning* (pp. 263–286). The Guilford Press.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, 2(4), 419-436.

Stuss, D. T., & Levine, B. (2002). Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology*, *53*(1), 401–433. https://doi.org/10.1146/annurev.psych.53.100901.135220

Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious Games - An overview*.

Tong, T., Chignell, M., & DeGuzman, C. A. (2021). Using a serious game to measure executive functioning: Response inhibition ability. *Applied*

Neuropsychology: Adult, *28*(6), 673–684. https://doi.org/10.1080/23279095.2019.1683561

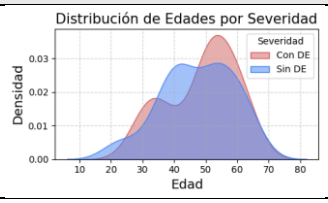
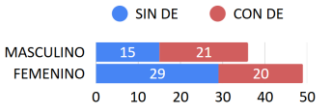
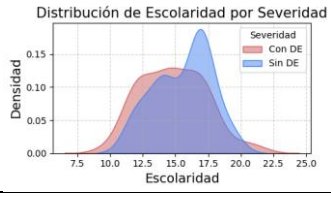
Tong, T., Chignell, M., Tierney, M. C., & Lee, J. (2016). A serious game for clinical assessment of cognitive status: Validation study. *JMIR Serious Games*, *4*(1), e7. https://doi.org/10.2196/games.5006

van Smeden, M., Moons, K. G. M., de Groot, J. A. H., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

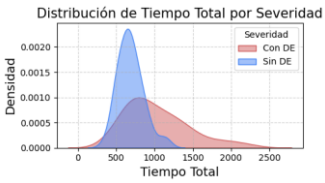
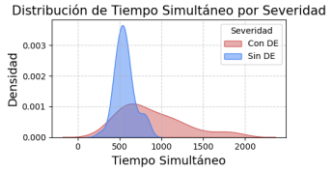
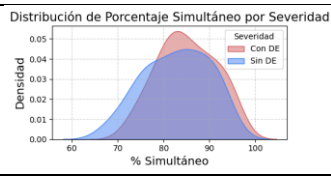
Varatharaj, A., Thomas, N., Ellul, M. A., Davies, N. W., Pollak, T. A., Tenorio, E. L., Sultan, M., Easton, A., Breen, G., Zandi, M., Coles, J. P., Manji, H., Al-Shahi Salman, R., Menon, D. K., Nicholson, T. R., Benjamin, L. A., Carson, A., Smith, C., Turner, M. R., & Solomon, T. (2020). Neurological and neuropsychiatric complications of COVID-19 in 153 patients: A UK-wide surveillance study. *The Lancet Psychiatry,* *7*(10), 875-882. https://doi.org/10.1016/S2215-0366(20)30287-X

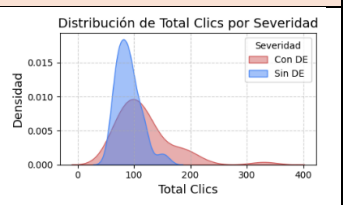
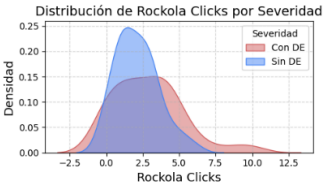
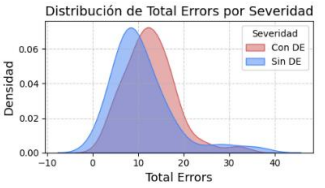
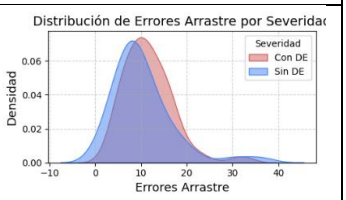
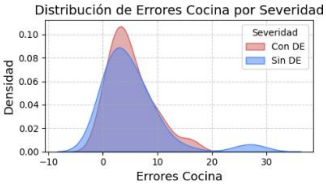
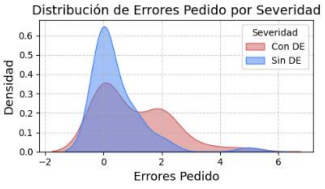
9. Anexos

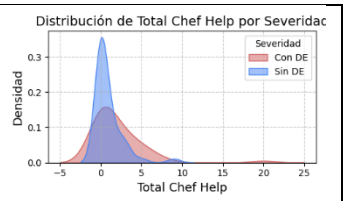
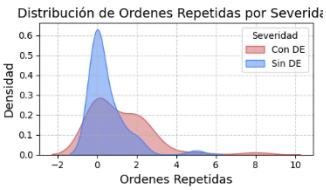
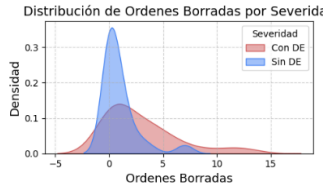
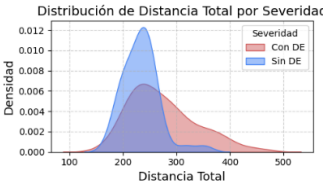
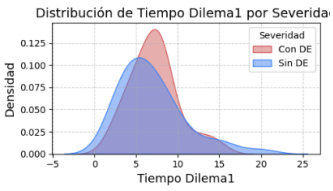
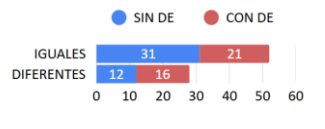
Anexo A: Análisis descriptivo y estadístico de Variables Demográficas

Variable	Tipo de Variable	Observaciones	Relación con Severidad	Gráfico
Edad (edad)	Discreta	Media Sin DE: 47.3 ± 11.3 Media Con DE:48.7 ± 10.9 Shapiro-Wilk < 0.05 U Mann Whitney > 0.05	Asociación **muy débil** y no estadísticamente significativa. $\rho = 0.047$ p-value > 0.05	
Sexo (sexo)	Categórica Binaria	El 57% del total de las participantes eran mujeres. Shapiro-Wilk < 0.05 U Mann Whitney > 0.05	Asociación **muy débil** y no estadísticamente significativa. $\chi^2 = 0.48$ p-value > 0.05 $\varphi = 0.074$	
Escolaridad (escolaridad)	Discreta	Media Sin DE: 15.6 ± 2.13 Media Con DE: 14.8 ± 2.46 Shapiro-Wilk < 0.05 U Mann Whitney > 0.05	Asociación **débil** y no estadísticamente significativa $\rho = -0.171$ p-value > 0.05	

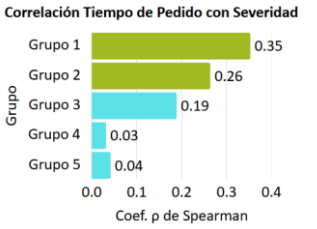
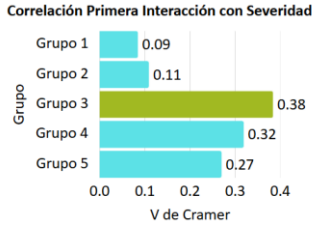
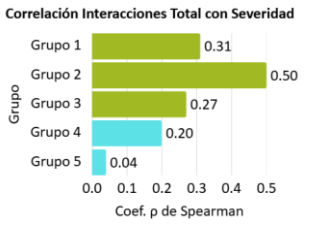
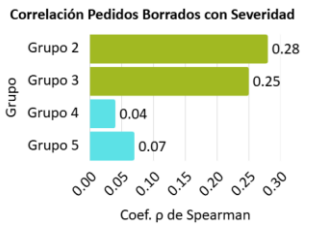
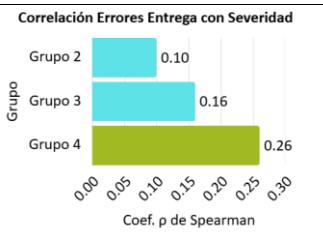
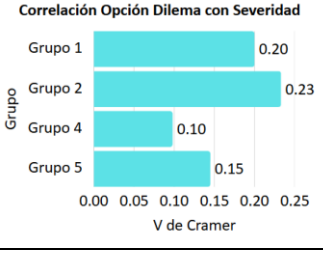
Anexo B: Análisis descriptivo y estadístico de Variables Manuales

Variable	Tipo de Variable	Observaciones	Relación con Severidad	Gráfico
Asociadas a Tiempo				
Tiempo total de juego (tiempo_total)	Continua	Tiempo total en segundos desde el inicio hasta el fin del juego. Promedio ≈ 856 s, con valores entre 360 y 2207 s	Asociación **moderada** y estadísticamente significativa $\rho = 0.488$ p-value < 0.05	
Tiempo simultáneo (tiempo_simultaneo)	Continua	Tiempo total en que más de un grupo estuvo activo simultáneamente. Promedio ≈ 721 s.	Asociación **fuerte** y estadísticamente significativa $\rho = 0.508$ p-value < 0.05	
Porcentaje simultáneo (porcentaje_simultaneo)	Continua	Porcentaje del tiempo total en que hubo simultaneidad de grupos. Promedio ≈ 83.9%.	Asociación **débil** y no estadísticamente significativa $\rho = 0.173$ p-value > 0.05	

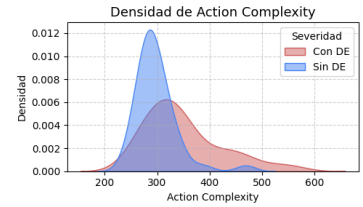
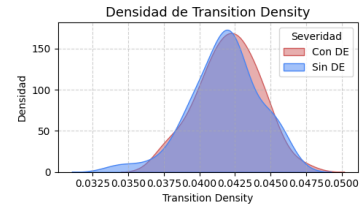
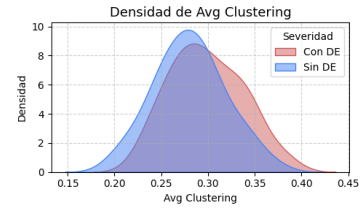
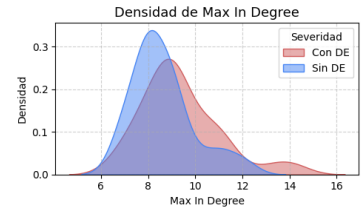
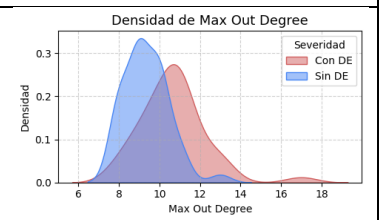
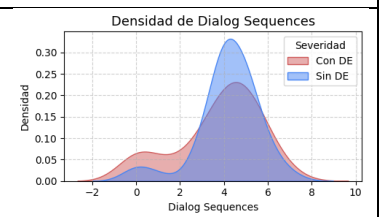
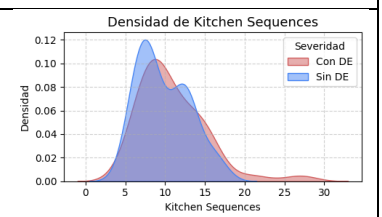
Asociadas a Clics				
Número total de Clics (total_clics)	Discreta	Total de interacciones realizadas mediante clics. Promedio ≈ 104	Asociación **moderada** y estadísticamente significativa $\rho = 0.343$ p-value < 0.05	
Clics sin efectos (clics_sin_efectos)	Discreta	Clics realizados que no provocaron ninguna acción o cambio en el juego. Promedio ≈ 31 clics	Asociación **débil** pero estadísticamente significativa $\rho = 0.241$ p-value < 0.05	
Clics en la rockola (rockola_clicks)	Discreta	Clics dirigidos a la rockola, que no tienen incidencia directa en la tarea.	Asociación **débil** y no estadísticamente significativa $\rho = 0.166$ p-value > 0.05	
Asociadas a Errores				
Errores totales (total_errors)	Discreta	Suma de todos los errores cometidos durante el juego. Incluye fallos en arrastre, cocina, entrega y pedidos.	Asociación **débil** pero estadísticamente significativa $\rho = 0.221$ p-value < 0.05	
Errores de arrastre (errores_arrastre)	Discreta	Fallos al intentar mover elementos incorrectamente	Asociación **débil** y no estadísticamente significativa $\rho = 0.18$ p-value > 0.05	
Errores en la cocina (errores_cocina)	Discreta	Fallos durante se habla con el Chef.	Asociación **muy débil** y no estadísticamente significativa $\rho = 0.051$ p-value > 0.05	
Errores en la entrega (errores_cocina)	Discreta	Entrega de pedidos erróneos, es decir, en la misma mesa o grupo, se le entrega al cliente equivocado.	Asociación **moderada** y estadísticamente significativa $\rho = 0.308$ p-value < 0.05	
Errores en los pedidos (errores_pedidos)	Discreta	Total de errores en la toma de pedidos.	Asociación **débil** pero estadísticamente significativa $\rho = 0.267$ p-value < 0.05	

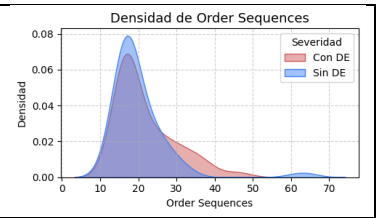
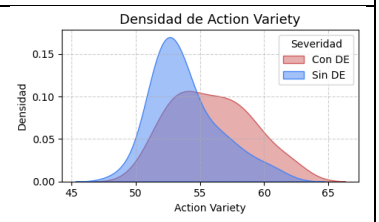
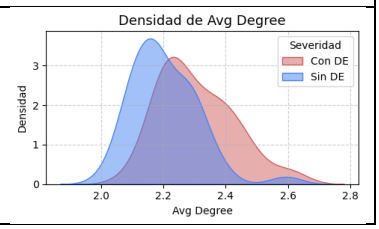
Ayudas del chef (total_chef_help)	Discreta	Cada vez que el jugador se equivoca en completar un pedido, el chef le avisa de que es el pedido equivocado.	Asociación **débil** pero estadísticamente significativa $\rho = 0.233$ p-value < 0.05	
Otras				
Órdenes repetidas (ordenes_repetidas)	Discreta	Cada vez que el jugador tuvo que pedirles a los clientes que repitieran el pedido.	Asociación **débil** pero estadísticamente significativa $\rho = 0.273$ p-value < 0.05	
Órdenes eliminadas (ordenes_borradas)	Discreta	Número de veces que se borraron órdenes activas	Asociación **moderada** y estadísticamente significativa $\rho = 0.366$ p-value < 0.05	
Distancia total recorrida (distancia_total)	Continua	Suma total del desplazamiento del jugador en el restaurante.	Asociación **moderada** y estadísticamente significativa $\rho = 0.351$ p-value < 0.05	
Dilemas 1 y 5				
Tiempo de respuesta dilema 1 (tiempo_respuesta_dilema1)	Continua	Tiempo (en segundos) que el jugador tardó en responder el primer dilema ético	Asociación **débil** y no estadísticamente significativa $\rho = 0.103$ p-value > 0.05	
Tiempo de respuesta dilema 5 (tiempo_respuesta_dilema5)	Continua	Tiempo de respuesta al dilema final, misma pregunta que el Dilema 1	Asociación **débil** y no estadísticamente significativa $\rho = 0.205$ p-value > 0.05	
Dilemas diferentes (dilemas_diferentes)	Binaria	Variable binaria: 1 si la decisión en el dilema 5 fue diferente a la del dilema 1. Se asocia con inconsistencia en valores	Asociación **muy débil** y no estadísticamente significativa. $\chi^2 = 0.48$ $p - valor > 0.05$ $\varphi = 0.074$	

Por cada grupo de clientes (G1-G5)

<table>
<tr><th>Variable</th><th>Tipo de Variable</th><th>Observaciones</th><th>Relación con Severidad</th><th>Gráfico</th></tr>
<tr><td>Tiempo de pedido (tiempo_toma_orden_gx)</td><td>Continua</td><td>Tiempo (en segundos) que el jugador tarda en tomar el pedido desde que selecciona la opción “Tomar Pedido” hasta que se envía a la Cocina.</td><td>Asociación moderada para el Grupo 1 y débil para el Grupo 2 ambas, estadísticamente significativas.
<table><tr><th>Grupo</th><th>p-value</th><th>ρ</th></tr><tr><td>G1</td><td><0.05</td><td>0.353</td></tr><tr><td>G2</td><td><0.05</td><td>0.264</td></tr><tr><td>G3</td><td>>0.05</td><td>0.189</td></tr><tr><td>G4</td><td>>0.05</td><td>0.032</td></tr><tr><td>G5</td><td>>0.05</td><td>0.042</td></tr></table></td><td> 
</td></tr>
<tr><td>Primera Interacción (type_label_gx)</td><td>Categórica</td><td>Primera opción que elige el participante cuando selecciona por primera vez un grupo de clientes.</td><td>Asociación moderada y estadísticamente significativa para el Grupo 3.
<table><tr><th>Grupo</th><th>χ^2</th><th>p-value</th><th>V de Cramer</th></tr><tr><td>G1</td><td>0.62</td><td>>0.05</td><td>0.085</td></tr><tr><td>G2</td><td>1.02</td><td>>0.05</td><td>0.109</td></tr><tr><td>G3</td><td>12.68</td><td><0.05</td><td>0.384</td></tr><tr><td>G4</td><td>8.75</td><td>>0.05</td><td>0.319</td></tr><tr><td>G5</td><td>6.28</td><td>>0.05</td><td>0.27</td></tr></table></td><td> 
</td></tr>
<tr><td>Interacciones totales (interacciones_gx)</td><td>Discreta</td><td>Número total de acciones realizadas con cada grupo de clientes</td><td>Asociación moderada para los grupos 1 y 2 y débil para el Grupo 3 todas, estadísticamente significativas.
<table><tr><th>Grupo</th><th>p-value</th><th>ρ</th></tr><tr><td>G1</td><td><0.05</td><td>0.31</td></tr><tr><td>G2</td><td><0.05</td><td>0.497</td></tr><tr><td>G3</td><td><0.05</td><td>0.271</td></tr><tr><td>G4</td><td>>0.05</td><td>0.203</td></tr><tr><td>G5</td><td>>0.05</td><td>0.037</td></tr></table></td><td> 
</td></tr>
<tr><td>Pedidos Borrados (ordenes_borradas_gx)</td><td>Discreta</td><td>Número de veces que se eliminó una orden antes de completarla.</td><td>Asociación Débil para los Grupos 2 y 3, ambas, estadísticamente significativa.
<table><tr><th>Grupo</th><th>p-value</th><th>ρ</th></tr><tr><td>G2</td><td><0.05</td><td>0.277</td></tr><tr><td>G3</td><td><0.05</td><td>0.248</td></tr><tr><td>G4</td><td>>0.05</td><td>0.037</td></tr><tr><td>G5</td><td>>0.05</td><td>0.065</td></tr></table></td><td> 
</td></tr>
<tr><td>Errores al entregar pedido (errores_entrega_gx)</td><td>Discreta</td><td>Fallos al entregar pedidos incorrectos a distinto comensal en cada grupo de clientes.</td><td>Asociación Débil y estadísticamente significativa para el Grupo 4.
<table><tr><th>Grupo</th><th>p-value</th><th>ρ</th></tr><tr><td>G2</td><td>>0.05</td><td>0.103</td></tr><tr><td>G3</td><td>>0.05</td><td>0.16</td></tr><tr><td>G4</td><td><0.05</td><td>0.255</td></tr></table></td><td> 
</td></tr>
<tr><td>Opción de Dilema (dilemma_option_gx)</td><td>Categórica</td><td>Opción escogida en cada uno de los dilemas propuestos. Para el Grupo 3, solo había una opción de respuesta, por lo que no fue considerado.</td><td>Asociación débil o muy débil para todos los grupos pero no estadísticamente significativa
<table><tr><th>Grupo</th><th>χ^2</th><th>p-value</th><th>V de Cramer</th></tr><tr><td>G1</td><td>3.42</td><td>>0.05</td><td>0.2</td></tr><tr><td>G2</td><td>4.67</td><td>>0.05</td><td>0.233</td></tr><tr><td>G4</td><td>0.83</td><td>>0.05</td><td>0.098</td></tr><tr><td>G5</td><td>1.8</td><td>>0.05</td><td>0.145</td></tr></table></td><td> 
</td></tr>
</table>

Anexo C: Análisis descriptivo y estadístico de Variables de Grafos

Variable	Tipo de Variable	Observaciones	Relación con Severidad	Gráfico
Complejidad de acciones (action_complexity)	Continua	Representa cuán estructuradas o desordenadas son las secuencias de acciones.	Asociación **moderada** y estadísticamente significativa $\rho = 0.423$ p-value < 0.05	
Densidad de transiciones (transition_density)	Continua	Mide qué tan densamente conectadas están las acciones entre sí en el grafo. Una alta densidad puede indicar que el jugador conecta casi todas las acciones posibles entre sí.	Asociación **muy débil** y no estadísticamente significativa $\rho = 0.066$ p-value > 0.05	
Promedio de clustering (avg_clustering)	Continua	Evalúa la tendencia de ciertas acciones a formar grupos cerrados de repeticiones (triángulos en el grafo).	Asociación **débil** y estadísticamente significativa $\rho = 0.234$ p-value < 0.05	
Máximo grado de entrada (max_in_degree)	Discreta	Indica cuántas acciones distintas apuntan a una misma acción específica. Puede sugerir que el jugador frecuentemente regresa a un mismo tipo de acción desde muchos puntos distintos.	Asociación **débil** y estadísticamente significativa $\rho = 0.269$ p-value < 0.05	
Máximo grado de salida (max_out_degree)	Discreta	Muestra desde qué acción el jugador se ramifica más hacia otras. Un valor alto sugiere que desde una acción específica, el jugador explora más posibilidades.	Asociación **moderada** y estadísticamente significativa $\rho = 0.47$ p-value < 0.05	
Secuencias de diálogos (dialog_sequences)	Discreta	Cuenta las cadenas de acciones relacionadas con el diálogo, con cada grupo de clientes.	Asociación **muy débil** y no estadísticamente significativa $\rho = -0.076$ p-value > 0.05	
Secuencias de cocina (kitchen_sequences)	Discreta	Mide la cantidad de ciclos de acciones en cocina, es decir, cuando entra a hablar con el chef.	Asociación **débil** y no estadísticamente significativa $\rho = 0.13$ p-value > 0.05	

Secuencias de órdenes (order_sequences)	Continua	Representa las cadenas de acciones desde que el jugador toma un pedido hasta que lo entrega	Asociación **débil** y no estadísticamente significativa $\rho = 0.114$ p-value > 0.05	
Variedad de acciones (action_variety)	Continua	Cuantifica cuántos tipos distintos de acciones realiza el jugador en la partida.	Asociación **moderada** y estadísticamente significativa $\rho = 0.345$ p-value < 0.05	
Grado promedio (avg_degree)	Continua	Promedio de conexiones por acción. Si es alto, el jugador tiende a relacionar cada acción con muchas otras.	Asociación **moderada** y estadísticamente significativa $\rho = 0.44$ p-value < 0.05	

Anexo D: Variables usadas de Pruebas Neuropsicológicas

Prueba Neuropsicológica	Variables	Tipo de Variable
DEX (Inventario Disfunción Ejecutiva)	total_dex_informante	Discreta
Dígitos (WAIS)	span_dd (Directo), ptje_e_dd (Directo Estándar), span_di (Inverso), ptje_e_di (Inverso Estándar)	Discreta / Continua
TMT (Trail Making Test)	puntaje_z_tmta, puntaje_z_tmtb	Continua
Torre de Londres (TOL)	pc_tol1 (respuestas correctas), pc_tol2 (movimientos), pc_tol3 (tiempo), pc_tol4 (violaciones de tiempo), pc_tol5 (violaciones de reglas)	Discreta
FAB (Frontal Assessment Battery)	tot_fab	Discreta
d2-R (atención visual)	conc_pc (concentración), vel_pc (velocidad), pre_pc (precisión)	Discreta
MoCA (Montreal Cognitive Assessment)	total_visuoespacial_moca, total_id_moca, total_atenyconc_moca, total_abstraccion_moca, total_mem_moca, puntaje_moca	Discreta

Anexo E: Escala de Dificultad Percibida (DP13-CL)

Le pedimos que evalúe la dificultad de las actividades que acaba de realizar en el Serious Game, utilizando para ello la siguiente escala. En primer lugar, intente situar su evaluación seleccionando algunas de las siete expresiones verbales, que abarcan un continuo desde "extremadamente fácil" hasta "Extremadamente Difícil". A continuación, puede refinar la respuesta mediante la escala numérica de 13 puntos, donde el primer nivel (1 punto) significa que la experiencia fue Extremadamente Fácil para usted. Por el contrario, el último nivel (13 puntos) corresponde a la dificultad máxima con la que puede evaluar la experiencia. Haga su evaluación de la forma más objetiva posible, evitando la sobreestimación o la subestimación.

Puntos	Categorías
1	Extremadamente Fácil
2	
3	Muy Fácil
4	
5	Fácil
6	
7	Ni Fácil Ni Difícil
8	
9	Difícil
10	
11	Muy Difícil
12	
13	Extremadamente Difícil

Anexo F: Cuestionario SUS (*System Usability Scale)*

Las respuestas se registran en una escala **Likert de 1 a 5** (1 = totalmente en desacuerdo, 5 = totalmente de acuerdo).

Ítem	Enunciado	Tipo
1	Creo que usaría este sistema frecuentemente	Positivo (+)
2	Encuentro este sistema innecesariamente complejo	Negativo (−)
3	Creo que el sistema fue fácil de usar	Positivo (+)
4	Creo que necesitaría ayuda de una persona con conocimientos técnicos para usar este sistema	Negativo (−)
5	Las funciones de este sistema están bien integradas	Positivo (+)
6	Creo que el sistema es muy inconsistente	Negativo (−)
7	Imagino que la mayoría de la gente aprendería a usar este sistema en forma muy rápida	Positivo (+)
8	Encuentro que el sistema es muy difícil de usar	Negativo (−)
9	Me siento confiado al usar este sistema	Positivo (+)
10	Necesité aprender muchas cosas antes de ser capaz de usar este sistema	Negativo (−)

Para calcular el puntaje asociado al cuestionario, se debe proceder de la siguiente manera:

- Para los ítems **positivos** (1, 3, 5, 7, 9): Se resta 1 al puntaje dado: $P - 1$
- Para los ítems **negativos** (2, 4, 6, 8, 10): Se resta el puntaje a 5: $5 - P$
- Sumar todos los resultados parciales: S
- Calcular el puntaje total SUS:

$$\text{Puntaje SUS} = S \cdot 2.5$$

El resultado estará en un rango de 0 a 100.

Anexo G: Script de Descarga de Datos desde la API de Neuronat

Este script en Python permite conectarse al servidor de Neuronat, autenticarse mediante una API REST y descargar los registros de sesiones de juego en formato JSON, organizándolos por código de participante (ID anónimo) y clasificándolos por carpetas locales. Este procedimiento fue esencial para construir el dataset del clasificador. Las credenciales y direcciones reales han sido ocultadas o modificadas por seguridad

```
import requests
import json
import os

# Obtener token de autenticación
def get_auth_token():
    auth_url = "http://<host>:<puerto>/api/authenticate"
    auth_data = {"username": "usuario", "password": "contraseña"}
    response = requests.post(auth_url, json=auth_data)
    return response.json().get('id_token')

# Obtener todos los registros con paginación
def get_all_data(auth_token):
    base_url = "http://<host>:<puerto>/api/neuronat-data"
    headers = {"Authorization": f"Bearer {auth_token}"}
    all_data, page = [], 0
    while True:
        response = requests.get(base_url, headers=headers, params={'page':
page})
        data = response.json()
        if not data:
            break
        all_data.extend(data)
        page += 1
    return all_data

# Limpiar nombres de carpetas
def clean_codigo(codigo):
    return ''.join(c for c in codigo if c.isalnum() or c in ['_', '-
']).strip().replace(' ', '_')

# Guardar cada JSON en su carpeta
def save_json_by_code(data_list):
    root_folder = 'datos_descargados'
    os.makedirs(root_folder, exist_ok=True)
    for item in data_list:
        try:
            codigo = clean_codigo(item.get('codigo', 'sin_codigo'))
```

```
            folder = os.path.join(root_folder, codigo)
            os.makedirs(folder, exist_ok=True)

            # Guardar archivo principal
            json_data = json.loads(item['json'])
            filename = os.path.join(folder, f"data_{item['id']}.json")
            with open(filename, 'w', encoding='utf-8') as f:
                json.dump(json_data, f, ensure_ascii=False, indent=4)

            # Guardar secciones si son listas
            for entry in json_data:
                for key, value in entry.items():
                    if isinstance(value, list):
                        section_path = os.path.join(folder,
f"{key}_{item['id']}.json")
                        with open(section_path, 'w', encoding='utf-8') as
f:
                            json.dump(value, f, ensure_ascii=False,
indent=4)
        except Exception as e:
            print(f"Error al procesar ID {item.get('id')}: {e}")

# Ejecución principal
token = get_auth_token()
if token:
    registros = get_all_data(token)
    save_json_by_code(registros)
```

Anexo H: Captura de la interfaz web/API de Neuronat

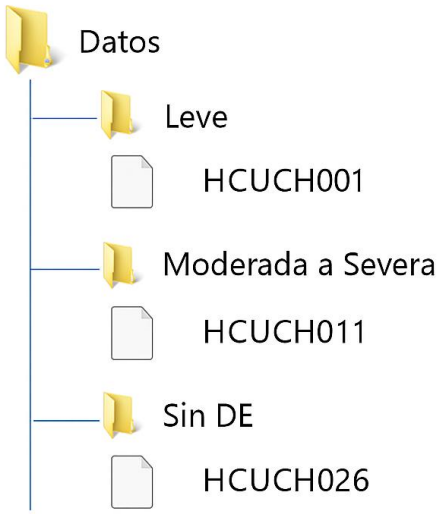
Anexo I: Ejemplo de estructura JSON obtenida del jugador


```
[{
    "ID": "HCUCH/HS/MS",
    "name": "Ramón Montenegro",
    "age": "65",
    "gender": "",
    "occupation": "Guillotinero",
    "character": "Hombre",
    "TimeStamp": "14-05-2025 13:07:16"
},{
    "listActions": [
        "LoadGroupG1##14-05-2025 13:07:40",
        "ClickTableG1-None##14-05-2025 13:07:43",
        "TakeOrder-SelectFood-jugo##14-05-2025 13:07:57",
        "ClickCompleteOrder##14-05-2025 13:07:57",
        "LoadGroupG2##14-05-2025 13:08:00",
        ... ]
},{
    "listKitchen": [{
            "table": 1,
            "foodCombo": "Combo 1",
            "timestamp": "14-05-2025 13:09:29"}]
},{
    "listMoving": [
        {
            "x": 0.45833301544189455,
            "y": 0.858333170413971,
            "timestamp": "14-05-2025 13:07:38"
        },
        {
            "x": -0.46666669845581057,
            "y": 0.6333333849906921,
            "timestamp": "14-05-2025 13:07:39"
        },
        ...    ]
},{
    "listInputs": [{
            "x": -2.4666666984558107,
            "y": 0.925000011920929,
            "timestamp": "14-05-2025 13:07:42"},
        ...   ]
},{
    "listGiveEntry": [{
            "food": 1,
            "character": "Hombre Solo",
            "timestamp": "14-05-2025 13:09:35"},
    ],
    "listTakeEntry": [{
            "listFood": [
                1,
                10],
            "timestamp": "14-05-2025 13:07:57"}
    ],
    "listProgressEntry": [{
            "progress": 1,
            "timestamp": "14-05-2025 13:07:57"},
        {
            "progress": 3,
            "timestamp": "14-05-2025 13:09:36"
        },
        {  "optionDilemma": 1,"timestampDilemma": "14-05-2025 13:09:03"
```

Anexo J: Estructura de Carpeta de Datos JSON

Los datos descargados desde la API de Neuronat fueron organizados localmente en una estructura de carpetas basada en la clasificación clínica de disfunción ejecutiva (DE), asignada por el juicio experto. Cada subcarpeta contiene archivos .json individuales por jugador, identificados mediante un código anónimo (ej. HCUCH001.json).

Esta organización fue para facilitar la gestión de muestras durante el proceso de preprocesamiento y entrenamiento del clasificador.



- Cada archivo contiene toda la información de la sesión en un único objeto JSON (acciones, dilemas, inputs, desplazamientos, etc.).
- La nomenclatura de los archivos (ej. HCUCH020.json) se utilizó como identificador único durante el análisis y clasificación.
- Los archivos fueron generados y almacenados tras un proceso automatizado vía API REST usando Python.

Anexo K: Funciones de Métricas

Debido a la longitud del código realizado para calcular las métricas a partir de los JSON, se menciona la lógica de estas funciones en este anexo.

```
def get_timestamp(action_string):
    """Extrae el timestamp de una acción 'Accion##dd-mm-YYYY
HH:MM:SS'."""
    try:
        return datetime.strptime(action_string.split("##")[1],
                                 "%d-%m-%Y %H:%M:%S")
    except (IndexError, ValueError):
        return None


def calcular_metricas_avanzadas(actions):
    """
    Calcula:
      • tiempo simultáneo entre mesas
      • % del tiempo total en modo simultáneo
      • tiempos de respuesta en dilemas 1 y 5
      • consistencia entre decisiones de dilema
    Retorna un dict con 6 métricas.
    """
    # ... implementación abreviada...
    return metrics


def calcular_distancia_total(data):
    """Distancia Euclídea recorrida excluyendo puntos duplicados."""
    # ... implementación abreviada ...


def extraer_variables_mesa(actions, grupo, enter_index=None):
    """
    Para cada grupo (G1-G5) extrae:
      - latencia hasta tomar pedido
      - tipo de interacción inicial (type_label)
      - errores de entrega y de pedido
      - total de interacciones
    """
    # ... implementación abreviada ...


def extract_game_metrics(data, severity, filename):
    """
    Orquesta:
      1. Métricas básicas (tiempos, errores globales)
      2. Métricas avanzadas (de la función anterior)
      3. Métricas por mesa G1-G5
      4. Métricas de grafo sobre la secuencia de acciones
    # ... implementación abreviada ...
```

Anexo L: Fragmento del Pipeline de Clasificación

Se documenta la comparación multitrial de modelos (Random Forest y XGBoost) sobre cinco datasets filtrados, la validación cruzada anidada y el entrenamiento final de los clasificadores para detección binaria de disfunción ejecutiva.

1. Librerías Usadas:

```
import pandas as pd
import numpy as np
from pathlib import Path
from sklearn.model_selection import StratifiedKFold, GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.feature_selection import VarianceThreshold, SelectKBest,
mutual_info_classif
from sklearn.metrics import (accuracy_score, precision_score, recall_score,
                             f1_score, roc_auc_score, confusion_matrix, make_scorer)
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
import joblib
```

2. Carga y selección de variables

```
# Datos preprocesados (ver Anexo A)
data = pd.read_csv("preprocessed_data.csv")

datasets = {
    "data_neuronat": data.drop([...variables socio-demo y neuropsi...], axis=1),
    "data_manual": data.drop([...solo métricas manuales...],       axis=1),
    "data_manual_graph":  data.drop([...métricas socio-demo...],         axis=1),
    "data_demo":          data[["severidad", "edad", "escolaridad", "sexo"]],
    "data_neuropsi":      data[["severidad", "dex", "span_dd", ...]]
}
```

3. Preprocesamiento Común

```
binary_cols      = ["dilemas_diferentes"]
categorical_cols = ["sexo", "primera_interaccion_g1", "dilemma_option_g1", ...]  #
solo si existen

def get_continuous(X):
    return [c for c in X.columns if c not in categorical_cols + binary_cols +
["severidad"]]

def build_preprocessor(X):
    return ColumnTransformer([
        ("num", StandardScaler(),            get_continuous(X)),
        ("cat", OneHotEncoder(handle_unknown="ignore"),
                        [c for c in categorical_cols if c in X.columns]),
        ("bin", "passthrough",               [c for c in binary_cols if c in
X.columns])
    ])
```

4. Selector determinístico por Información Mutua

```
def mutual_info_fixed(X, y):
    return mutual_info_classif(X, y, random_state=42)
```

5. Modelos y rejillas de búsqueda

```
models_grids = {
    "Random Forest": {
        "model": RandomForestClassifier(random_state=42, class_weight="balanced"),
        "params": {
            "model__n_estimators":     [100, 200, 300],
            "model__max_depth":        [3, 5, 10, None],
            "model__min_samples_split":[2, 5, 10],
            "model__min_samples_leaf": [1, 2, 4],
            "model__max_features":     ["sqrt", "log2", None]
        }
    },
    "XGBoost": {
        "model": XGBClassifier(use_label_encoder=False,
                               eval_metric="logloss",
                               random_state=42,
                               deterministic=True,
                               nthread=1),
        "params": {
            "model__n_estimators": [100, 200],
            "model__max_depth":    [3, 5, 10],
            "model__learning_rate":[0.01, 0.1]
        }
    }
}
```

6. Validación cruzada anidada multitrial

```
outer_cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
inner_cv = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)
NUM_TRIALS = 10
results = []

for trial in range(1, NUM_TRIALS + 1):
    for name, df in datasets.items():
        X, y = df.drop("severidad", axis=1), df["severidad"]
        prep = preprocessor(X)

        for model_name, cfg in models_grids.items():
            pipe = Pipeline([
                ("preprocessor", prep),
                ("variance_filter", VarianceThreshold(0.02)),
                ("select", SelectKBest(mutual_info_fixed, k=20)),
                ("model", cfg["model"])
            ])

            nested = GridSearchCV(pipe,
                                  param_grid=cfg["params"],
                                  cv=inner_cv,
                                  scoring="recall",
                                  n_jobs=-1)
```

```
            fold = 0
            for train_idx, test_idx in outer_cv.split(X, y):
                fold += 1
                nested.fit(X.iloc[train_idx], y.iloc[train_idx])

                y_pred  = nested.best_estimator_.predict(X.iloc[test_idx])
                y_prob  = nested.best_estimator_.predict_proba(X.iloc[test_idx])[:,
1]

                results.append({
                    "trial":      trial,
                    "dataset":    name,
                    "model":      model_name,
                    "fold":       fold,
                    "recall":     recall_score(y.iloc[test_idx], y_pred),
                    "precision":  precision_score(y.iloc[test_idx], y_pred),
                    "specificity": specificity_score(y.iloc[test_idx], y_pred),
                    "roc_auc":    roc_auc_score(y.iloc[test_idx], y_prob),
                    "best_params": nested.best_params_
                })
```

El bucle genera 50 evaluaciones por dataset (10 trials × 5 folds) con optimización interna de hiperparámetros.

7. Entrenamiento y guardado de los modelos finales

```
for name, df in datasets.items():
    if name == "data_neuropsi":      # no se entrena RF final en neuropsi
        continue

    X, y = df.drop("severidad", axis=1), df["severidad"]
    prep = preprocessor(X)

    final_rf = RandomForestClassifier(
        n_estimators=100, max_depth=3,
        min_samples_split=2, min_samples_leaf=1,
        max_features="sqrt", random_state=42,
        class_weight="balanced"
    )

    final_pipe = Pipeline([
        ("preprocessor", prep),
        ("variance_filter", VarianceThreshold(0.02)),
        ("model", final_rf)
    ])

    final_pipe.fit(X, y)
    joblib.dump(final_pipe,
f"modelos_entrenados_rf/{name}_RandomForest_final.joblib")
```

8. Notas de Reproducibilidad
   - Python 3.10, scikit-learn 1.5, xgboost 2.0
   - Semillas fijadas (np.random.seed(11) y random_state=42)
   - Duración total ≈ 65–70 min en un Ryzen 7 5800X / 32 GB RAM