UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

EXPLORACIÓN DE DATASETS USANDO IA GENERATIVA VS EXPERTO HUMANO EN LENGUAJE R

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

JORGE FRED ALVAREZ SALAYA

BAJO LA DIRECCIÓN DE:

DRA. JUANA CANUL REICH

CUNDUACÁN, TABASCO, A: DICIEMBRE 2025

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

EXPLORACIÓN DE DATASETS USANDO IA GENERATIVA VS EXPERTO HUMANO EN LENGUAJE R

TESIS PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

JORGE FRED ALVAREZ SALAYA

BAJO LA DIRECCIÓN DE:

DRA. JUANA CANUL REICH

CUNDUACÁN, TABASCO, A: DICIEMBRE 2025

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día tres del mes de Diciembre del año 2025, el que suscribe **Jorge Fred Alvarez Salaya**, alumno del Programa de la **Maestría en Ciencias de la Computación** con número de matrícula **232H21008**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Maestría y titulada **Exploración de datasets usando IA Generativa vs experto humano en lenguaje R**, dirigida por la Dra. Juana Canul Reich .

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 03 de Diciembre de 2025.

Estudiante: Jorge Fred Alvarez Salaya

UJAT
UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
"ESTUDIO EN LA DUDA, ACCIÓN EN LA FE"

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

2025 AÑO DE LA Mujer Indígena

Cunduacán, Tabasco, a 03 de diciembre de 2025
Oficio No. 2456 /2025/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Jorge Fred Álvarez Salaya
Egresado de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Exploración de Datasets usando IA Generativa vs. Experto Humano en Lenguaje R"***, para presentar examen y obtener el Grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

Dr. Óscar Alberto González González
Director

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Mtra. Yenny Lorena Dussán Rojas. – Encargada del despacho de la Coordinación de Posgrado.
Archivo.
Consecutivo.

DR.*OAGG/YLDR

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690.
Cunduacán, Tabasco, México.
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Carta de Cesión de Derechos

Villahermosa, Tabasco a 03 de Diciembre de 2025.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Exploración de datasets usando IA Generativa vs experto humano en lenguaje R**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Jorge Fred Alvarez Salaya

TESTIGOS

Dra. Juana Canul Reich

Dr. Oscar Alberto Chávez Bosquez

Dedicatoria

En primer lugar, a Dios, por ser mi guía, mi fortaleza en los momentos de incertidumbre y por permitirme llegar hasta este punto con salud y sabiduría.

A las dos mujeres que han sido el pilar de mi vida: mi madre, Hermila Salaya, y mi abuela, Irma Pablo Adorno. Gracias infinitas a ambas por su amor incondicional, por sus sacrificios y por haberme apoyado tanto en cada paso de este camino. Este logro es tan suyo como mío.

De manera muy especial, dedico esto a mi sobrinita, Arantza. Sé que ahora eres pequeña, pero deseo que cuando crezcas y encuentres este documento, sepas que fuiste una motivación inmensa para mí. Espero que esto te sirva de ejemplo para saber que los sueños se cumplen con esfuerzo.

Expreso mi más sincero agradecimiento a la Dra. Juana Canul Reich, directora de esta tesis, por guiarme con paciencia en el camino del conocimiento y por su invaluable mentoría académica. Asimismo, extiendo mi gratitud a los doctores Oscar Alberto Chávez Bosquez, Cristina López Ramírez y José Adán Hernández Nolasco, cuyas observaciones y retroalimentación crítica fueron fundamentales para fortalecer este documento.

Finalmente, a mis compañeros y amigos de maestría: Oscar Fabian, Luis Ramon Tercero, Orlando Flores y Jesus Manuel. Gracias por el intercambio de ideas y el apoyo mutuo que enriqueció mi aprendizaje.

Índice general

Índice de figuras

Índice de tablas

Resumen

Este trabajo de investigación analiza el papel de la inteligencia artificial generativa en la ciencia de datos, comparando el desempeño de GPT-4o con el de un experto humano a lo largo del ciclo de vida CRISP-DM. El trabajo se inscribe en un contexto donde el volumen de datos y la presión por automatizar tareas analíticas hacen cada vez más relevante la posibilidad de delegar parte del proceso a modelos de lenguaje de gran tamaño, sin perder rigor ni calidad en problemas reales del ámbito de la salud.

La metodología se basa en tres conjuntos de datos médicos (Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set), en los que se ejecuta el mismo flujo de trabajo por duplicado: por un lado, mediante scripts en R desarrollados por un experto, y por otro, a partir de indicaciones dirigidas a GPT-4o, que generan documentos R Markdown reproducibles. En ambos casos se entrenan los mismos algoritmos de aprendizaje automático y se comparan sus resultados con métricas estándar.

Los resultados muestran que GPT-4o puede igualar o incluso superar al humano en problemas con estructura más clara (autismo y epilepsia), pero ofrece un rendimiento sensiblemente inferior y más inestable en el caso de diabetes. A partir de ello, la tesis concluye que la IA generativa no sustituye al experto, aunque sí puede funcionar como asistente que automatiza tareas repetitivas y acelera la experimentación, dentro de esquemas híbridos donde el control y la decisión final siguen recayendo en la persona especialista.

Palabras clave: *IA generativa; ciencia de datos; CRISP-DM; GPT-4o; agentes inteligentes.*

Abstract

This study analyzes the role of generative artificial intelligence in data science by comparing the performance of GPT-4o with that of a human expert across the CRISP-DM lifecycle. The study is set within a context of growing data volumes and increasing pressure to automate analytical tasks, which makes it increasingly relevant to delegate parts of the process to large language models without sacrificing rigor or quality in real-world, health-related problems.

The methodology is based on three medical datasets (Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set, and Diabetes Data Set), in which the same workflow is executed twice: on the one hand, through R scripts developed by an expert, and on the other, through instructions given to GPT-4o, which produces reproducible R Markdown documents. In both approaches, the same machine learning algorithms are trained and their results are compared using standard metrics.

The findings show that GPT-4o can match or even surpass human performance in problems with a clearer structure (autism and epilepsy), but delivers noticeably lower and more unstable performance in the diabetes case. Based on this, the thesis concludes that generative AI does not replace the expert, although it can function as an assistant that automates repetitive tasks and accelerates experimentation, within hybrid schemes in which control and final decision-making remain in the hands of the human specialist.

***Keywords**: generative AI; data science; CRISP-DM; GPT-4o; intelligent agents.*

Capítulo 1

Generalidades

1.1. Introducción

En las décadas previas a la revolución digital, la investigación y el análisis de datos se realizaban de manera manual o con herramientas computacionales básicas que eran limitadas en su capacidad de procesamiento. La llegada de Internet marcó un cambio radical, abriendo caminos para una comunicación global instantánea y el acceso a un vasto océano de información. Este cambio no solo transformó la manera en que interactuamos y compartimos conocimiento, sino que también facilitó el desarrollo y la expansión de la ciencia de datos (Rodríguez Yunta, 2017; Sarker, 2021; Syafganti, 2018).

Con la expansión de internet, los datos comenzaron a generarse a una escala sin precedentes, lo que llevó a la necesidad de desarrollar métodos más sofisticados para lograr su análisis y gestión. La ciencia de datos emergió como un campo interdisciplinario que combina elementos de estadística, análisis de datos y computación para extraer conocimientos y patrones significativos de grandes conjuntos de datos. Este desarrollo abrió nuevas posibilidades en campos tan variados como la medicina, la economía y la ingeniería (Milligan, 2022).

Paralelamente, la inteligencia artificial (IA) comenzó a ganar terreno, transformando radicalmente diversos aspectos de la sociedad. Desde sistemas de recomendación en plataformas de streaming hasta algoritmos de diagnóstico médico, la IA se integró en la vida cotidiana, demostrando su capacidad para mejorar y acelerar procesos tradicionales

(Lee, 2020).

La invención de la IA generativa marcó un cambio de paradigma significativo. Estos sistemas, capaces de generar contenido nuevo y realizar tareas complejas de procesamiento de datos, ofrecen mejoras en eficiencia y un aumento en la capacidad analítica.

La IA generativa, ejemplificada por modelos como los modelos GPT, desarrollados por OpenAI, han demostrado ser una herramienta poderosa, acelerando y enriqueciendo la forma de realizar investigaciones y análisis de datos. En el contexto de la ciencia de datos, esta tecnología ofrece una alternativa novedosa a los métodos tradicionales de programación y análisis. Históricamente, la programación ha evolucionado desde lenguajes de bajo nivel como el ensamblador, pasando por lenguajes de medio nivel como C, hasta llegar a lenguajes de alto nivel como Python y R. La IA generativa representa el siguiente paso en esta evolución, proporcionando una forma más intuitiva y eficiente de interactuar con los datos y desarrollar soluciones (Aljanabi et al., 2023).

Es importante destacar que la llegada de la IA generativa no implica la sustitución de los trabajos existentes, sino una transformación en la manera en que estos se abordan. De forma similar a como los lenguajes de programación de alto nivel hicieron la programación más accesible y versátil, la IA generativa promete democratizar aún más el acceso al análisis de datos avanzado. Esto permite que usuarios de diversos niveles de experiencia técnica puedan aprovechar el poder de los algoritmos de IA para sus propios fines.

La propuesta de este trabajo de investigación tiene como objetivo comparar la capacidad que tiene la IA generativa, en particular, el modelo GPT-4o para realizar las etapas del ciclo de vida CRISP-DM para ciencia de datos. Se busca comparar la ejecución de cada etapa por parte de una IA generativa frente a la codificación manual de un experto humano. Se busca identificar las diferencias en cada etapa entre estos dos enfoques.

1.2. Planteamiento del problema

1.2.1. Definición del problema

La ciencia de datos, disciplina en constante evolución, se ha convertido en el núcleo principal del procesamiento de información, transitando desde sus orígenes en estadística e informática hasta convertirse en una disciplina independiente (Karthika et al., 2022; Urs y Minhaj, 2023). Con el desarrollo de la inteligencia artificial, algoritmos de Machine Learning, almacenamiento en la nube y el internet de las cosas, la ciencia de datos ha experimentado cambios de paradigma necesarios para poder abordar nuevos problemas y desafíos que las disciplinas tradicionales no estaban preparadas para abordar (Govindarajan, 2020; Palazhchenko et al., 2023; Zhang et al., 2022).

En la era del Big Data, la exploración de datos se ha convertido en una fase crítica para descubrir hallazgos significativos, pero de igual forma resulta ser intensiva en tiempo y recursos (Gracla et al., 2022; «Human Uncertainty and Ranking Error – The Secret of Successful Evaluation in Predictive Data Mining», 2017; Hüsing, 2021; Ma et al., 2023). Aquí es donde la Inteligencia Artificial (IA) generativa, y en particular los Modelos de Lenguaje de Gran Escala (LLMs) como los modelos GPT (Generative Pre-trained Transformer) desarrollados por OpenAI, emergen como innovadores. Estos modelos, con sus capacidades avanzadas de procesamiento de lenguaje natural, presentan un gran potencial para automatizar y optimizar tareas de ciencia de datos (Zhao et al., 2023). Sin embargo, la evaluación comparativa de su eficacia frente a los métodos convencionales sigue siendo una área poco explorada, principalmente debido a que estas tecnologías, como el modelo GPT-4o, son aún recientes y solo se han puesto a disposición del público en 2023, limitando así las oportunidades para su aplicación práctica y estudio en profundidad (Stadlmann y Zehetner, 2021).

Los métodos tradicionales de ciencia de datos, que requieren el uso intensivo de programación en lenguajes como R y Python como los más populares, aunque son efectivos, consumen una cantidad considerable de tiempo y recursos. De igual forma implementar estas técnicas requieren un alto grado de especialización en el conocimiento de esas

herramientas (Al-Haija, 2022). Por otro lado, los modelos de IA generativa como GPT-4o ofrecen un enfoque alternativo que puede simplificar estos procesos mediante el uso de prompts, proponiendo un análisis de datos más eficiente y creativo (Morris, 2023). En este estudio se propone examinar esta posibilidad, realizando una comparación detallada y técnica de la eficacia de los métodos tradicionales frente a soluciones basadas en IA generativa.

Específicamente, se evaluará cómo el modelo GPT-4o, en conjunto con herramientas como Noteable, maneja las tareas de ciencia de datos, desde la limpieza de los datos hasta el modelado predictivo y la visualización. Se emplearán métricas específicas y pruebas estadísticas para medir la precisión, eficiencia y capacidad de descubrimiento de hallazgos significativos en conjuntos de datos. Además, se discutirán las limitaciones de este enfoque, posibles sesgos en los modelos de IA, y cómo estos pueden afectar los resultados.

Los resultados de este estudio no solo buscarán evaluar la eficacia comparativa de la IA en la ciencia de datos, sino también su aplicabilidad práctica y su impacto en el campo de la computación. Este análisis podría ofrecer perspectivas valiosas sobre el papel emergente de la IA en la ciencia de datos, estableciendo un marco para futuras investigaciones y aplicaciones prácticas en un campo que se encuentra en constante desarrollo.

1.2.2. Delimitación de la investigación

Alcances

- *En esta investigación se buscará comparar el rendimiento de la IA generativa, específicamente modelos como GPT-4o, con métodos tradicionales de análisis en ciencia de datos.*
- *Se utilizarán los datasets Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set obtenidos del repositorio de datos públicos UCI Machine Learning para evaluar estos métodos, proporcionando hallazgos valiosos en su aplicación.*

- *Se entrenarán y compararán modelos predictivos utilizando los siguientes algoritmos de Machine Learning: Decision Tree, Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Neural Network, Random Forest, Boosted Tree y un Ensamble Heterogéneo.*
- *Las métricas utilizadas para medir el rendimiento de los modelos creados para ambos enfoques serán precisión, accuracy balanceado, sensitividad y especificidad.*
- *Se complementará de que formas la IA generativa puede mejorar las metodologías tradicionales en el análisis y manejo de datos.*

Limitaciones

- *Los resultados y conclusiones estarán limitados a los datasets específicos utilizados.*
- *Esta investigación se centra en utilizar el modelo GPT-4o desarrollado por OpenAI que se será utilizado para los experimentos.*

1.3. Pregunta de investigación

- *¿Puede un LLM como GPT-4o, realizar la etapa de conocimiento de datos del ciclo CRISP-DM con resultados comparables a los obtenidos por el código escrito por un experto humano?*
- *¿Puede un LLM como GPT-4o, realizar la etapa de preparación de datos del ciclo CRISP-DM con resultados comparables a los obtenidos por el código escrito por un experto humano?*
- *¿Puede un LLM como GPT-4o, alcanzar o superar los niveles de precisión logrados por el código escrito por un experto humano en la creación de modelos predictivos?*

1.4. Hipótesis

En experimentos paralelos de desarrollo de modelos predictivos, los modelos generados por un experto humano obtendrán un accuracy balanceado superior en al menos un 3 % respecto a los modelos generados por una IA generativa (GPT-4o), manteniendo niveles equivalentes de estabilidad en sensibilidad y especificidad.

1.5. Objetivos

Objetivo general *Evaluar el desempeño del modelo GPT-4o en la ejecución del ciclo de vida CRISP-DM a partir de prompts, en comparación con el correspondiente a la ejecución del código escrito por un experto humano.*

Objetivos específicos

a. *Realizar el análisis exploratorio de datos (EDA) para los conjuntos de datos Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set, identificando patrones, anomalías y relaciones, tanto a través de prompts con el modelo GPT-4o como mediante código escrito por un experto humano en LenguajeR.*

b. *Limpiar y preparar los datasets seleccionados aplicando técnicas de preprocesamiento de datos para tratar valores faltantes y atípicos, ejecutando los procedimientos tanto mediante scripts escritos en lenguaje R por un experto como a través de la generación de código con GPT-4o.*

c. *Desarrollar modelos predictivos para cada uno de los datasets mencionados, empleando algoritmos de machine learning (tales como Decision Tree, Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Neural Network, Random Forest, Boosted Tree y un Ensamble Heterogéneo), bajo ambos enfoques: experto humano y asistencia de IA generativa.*

d. *Obtener y comparar las métricas de rendimiento (precisión, accuracy balanceado, sensibilidad y especificidad) de los modelos generados por el experto humano frente a los producidos por GPT-4o, para determinar la eficacia y fiabilidad de la IA en tareas de ciencia de datos.*

1.6. Justificación

Hoy en día, los avances en el desarrollo de modelos de IA generativa han mostrado grandes avances en muchos campos y disciplinas, y en el campo de ciencia de datos, se han generado grandes expectativas. Estos modelos han demostrado su capacidad para realizar tareas complejas que antes estaban reservadas exclusivamente para analistas y programadores expertos. La habilidad que ha demostrado para abordar y resolver problemas complejos, genera una inquietud fundamental entre los profesionales de datos: determinar el alcance que tiene la IA generativa en la resolución de las diferentes etapas del proceso de análisis de datos y creación de modelos

De forma tradicional, un programador o analista de datos es quien escribe el código y maneja los procesos que abarcan desde la comprensión del negocio hasta la evaluación y el despliegue de las etapas de métodologías para ciencia de datos como CRISP-DM. Sin embargo, con la llegada de la IA generativa, como el modelo GPT-4o desarrollado por OpenAI, se ha presenciado un gran potencial. No solo ha mostrado ser un competidor para resolver tareas de este campo, sino también un posible aliado valioso. La confianza en sus capacidades y resultados son aspectos claves que se están explorando activamente. Es importante destacar que los resultados producidos por la IA generativa dependen en gran medida de la calidad y la precisión de los prompts que se le proporcionan. Esta dependencia subraya la necesidad de un entendimiento claro y una formulación precisa de los problemas a ser abordados. En este sentido, GPT-4o y modelos similares no vienen a reemplazar el trabajo humano, sino a complementarlo, ofreciendo nuevas herramientas y posibilidades para mejorar y acelerar los procesos de ciencia de datos.

Este proyecto representa un esfuerzo significativo para avanzar en el campo de la inteligencia artificial y el análisis de datos al centrarnos en analizar una gama diversa

de datasets, que incluyen áreas como la medicina. Se han elegido deliberadamente tres conjuntos de datos del ámbito de la salud (Autistic Spectrum Disorder, Epileptic Seizure Recognition y Diabetes) que presentan características estructurales contrastantes. Esta heterogeneidad, que abarca desde muestras pequeñas con valores faltantes y clases balanceadas, hasta conjuntos de alta dimensionalidad con miles de registros y fuerte desbalance de clases, es fundamental para el estudio. Permite evaluar no solo la capacidad de ejecución de la IA generativa, sino su adaptabilidad y criterio frente a los desafíos más comunes y dispares de la ciencia de datos, determinando así en qué escenarios específicos su desempeño se equipara, supera o diverge del experto humano.

Desde un punto de vista computacional, también se busca evaluar y comparar el rendimiento de la IA generativa frente a los métodos tradicionales. Esta investigación no solo aportará resultados sobre la confiabilidad de las capacidades actuales de las tecnologías de la IA generativa, sino que también servirá como una base para futuras mejoras y desarrollos en el campo del análisis de datos. La adaptabilidad y la innovación que caracterizan a la IA generativa podrían encontrar nuevas formas de procesamiento y análisis de datos, abriendo camino a soluciones más eficientes y efectivas (Jeha et al., 2023).

En lo referente a beneficios sociales, la aplicación de estas tecnologías en diversos sectores tiene el potencial de revolucionar la manera en que abordamos problemas críticos. En el sector de la salud, por ejemplo, la mejora en diagnósticos y tratamientos podría significar una atención médica más precisa y rápida, impactando positivamente en la vida de las personas (Jeha et al., 2023). Al no centrar la investigación únicamente en una sola área, se aspira a encontrar hallazgos significativos que podrían mejorar servicios, contribuyendo así a una mejor calidad de vida.

Con respecto a los beneficios económicos de este proyecto, se busca aportar conocimiento significativo al optimizar los procesos de la ciencia de datos. Se puede esperar una reducción de costos operativos y un aumento en la eficiencia de las tareas de análisis y procesamiento de datos. Esta mejora en la gestión de datos y en la toma de decisiones basada en análisis precisos podría traducirse en una mayor rentabilidad y competitividad en varias industrias.

Según el Banco Mundial, la IA generativa ha estado desarrollándose rápidamente

y ha atraído una atención significativa en los últimos años, con numerosos avances y descubrimientos. Se espera que el mercado de la IA generativa crezca rápidamente, con una tasa de crecimiento anual compuesta de 34.9 por ciento, alcanzando los 6.5 mil millones de dólares para el 2026 (World Bank, 2023).

Esto promete que este tipo de tecnologías continúe desarrollándose (Zohny et al., 2023). Se están realizando esfuerzos para abordar las preocupaciones sobre la precisión y confiabilidad de los resultados presentados por una IA generativa, incluida la IA explicable y las prácticas responsables de IA (Tang et al., 2023). A pesar de ser una tecnología relativamente nueva, las aplicaciones generativas de IA como ChatGPT ya se están utilizando en diversos campos, por lo que es esencial que los líderes en innovación se familiaricen con estas herramientas (Feuerriegel et al., 2023).

Existen muchos desafíos y temas controversiales con respecto al uso de la IA generativa, pero a pesar de esto, las aplicaciones generativas de IA como ChatGPT ya están siendo ampliamente adoptadas y aplicadas en diversos campos (Samuelson, 2023). Por lo tanto, es claro que la IA generativa ha venido para quedarse y seguirá utilizándose de formas innovadoras.

1.7. Metodología

Diseño de Investigación

Este estudio sigue un enfoque comparativo para evaluar el rendimiento de la IA generativa representada por el modelos como GPT-4o de OpenAI, frente a los métodos tradicionales de ciencia de datos. Se centrará en el uso de los datasets: Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set

La Figura 1.1 presenta el flujo de trabajo que ilustra el método comparativo que se utilizará en el proyecto para evaluar el rendimiento de los modelos de ciencia de datos creados con dos enfoques distintos: El enfoque tradicional, el cual es escrito utilizando

lenguaje R por un experto humano y el enfoque de IA generativa, utilizando GPT-4o. Cabe destacar que el método será usado para cada uno de los datasets definidos en este proyecto. A continuación, se describe cada etapa del método:

Etapa 1: Conocimiento de los datos *Se muestra la exploración inicial del dataset. Esta etapa es crucial para obtener una comprensión profunda de los datos. A continuación se detallan las actividades involucradas en esta fase:*

La Tabla 1.1 compara las actividades a realizar en esta etapa según cada enfoque.

Enfoque	Descripción
Experto humano	Utilizando el lenguaje de programación Lenguaje R, se llevan a cabo los procesos mencionados anteriormente mediante la codificación de los scripts específicos, sumado las librerías en R que facilitan el análisis exhaustivo y detallado.
IA Generativa	Con GPT-4o, el análisis será mediante prompts, solicitando identificar tendencias y, posiblemente, sugerir correlaciones relevantes. El conocimiento de datos a través de GPT-4o puede ser menos técnico y más accesible para aquellos que no están versados en programación estadística.

Tabla 1.1. Descripción de actividades en la etapa de conocimiento de los datos

Etapa 2: Preparación de los datos *En esta etapa los datos crudos se transforman realizando técnicas de limpieza de datos y preprocesamiento de datos para tratar valores faltantes, eliminación o corrección de datos duplicados o con errores, manejo de valores atípicos, normalización o estandarización de datos, codificación de variables categóricas, reducción de dimensionalidad. En la Tabla 1.2 se muestran la descripción de actividades de la etapa dos.*

Etapa 3: Modelado *Esta etapa se centra en construir modelos utilizando técnicas de machine learning que posteriormente serán evaluados por sus métricas de rendimiento. Esto implica definir los parámetros y los hiperparámetros que necesitan ser ajustados. En la Tabla 1.3 se muestran las actividades que se realizaran por cada enfoque.*

Enfoque	Descripción
Experto humano	Utilizando R, se ejecutan scripts para limpiar y preprocesar los datos, permitiendo un control detallado sobre cada paso del proceso.
IA Generativa	Se realizarán las tareas de limpieza y preprocesamiento a través de prompts específicos. Jupyter Notebooks será utilizado para registrar, replicar y ejecutar el código generado por GPT-4o.

Tabla 1.2. Descripción de actividades en la etapa de preparación de los datos

Enfoque	Descripción
Experto humano	Utilizando R, se ejecutan scripts para entrenar y ajustar los modelos, permitiendo un control detallado sobre cada paso del proceso.
IA Generativa	Se generarán los modelos a través de prompts específicos. Jupyter Notebooks será utilizado para registrar, replicar y ejecutar el código producido por GPT-4o.

Tabla 1.3. Descripción de actividades en la etapa de modelado.

Etapa 4: Evaluación de resultados *Se evaluarán los modelos generados usando un conjunto de métricas para cada enfoque. Las métricas específicas no se mencionan en el diagrama, las cuales son precisión, accuracy balanceado, sensitividad, especificidad .*

Etapa 5: Comunicación de resultados *Cada enfoque genera un reporte de resultados que resumirá los hallazgos y métricas obtenidos durante la evaluación de los modelos. Estos reportes son vitales para la interpretación y el análisis de los resultados.*

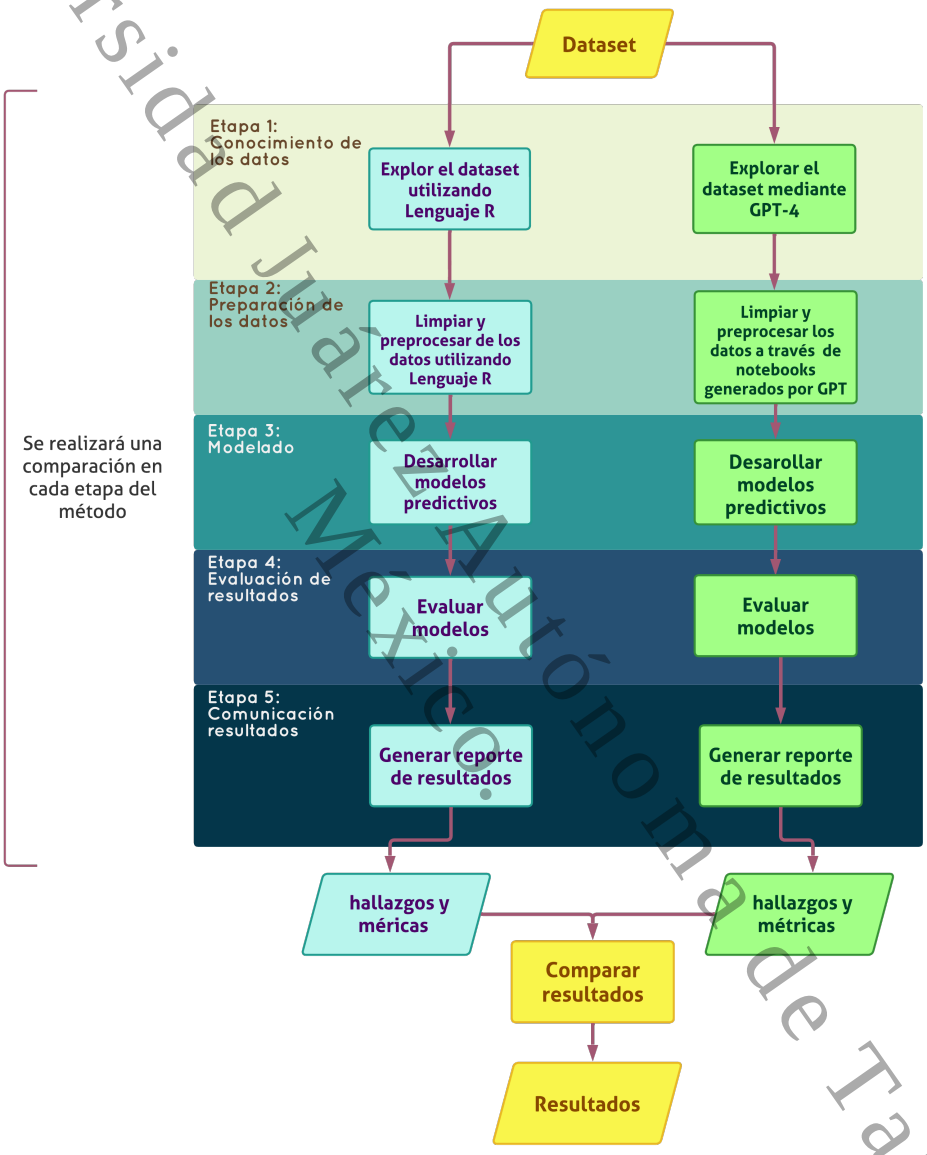

Figura 1.1. Diagrama del método

Capítulo 2

Fundamentos

2.1. Conceptos y teorías fundamentales de la investigación

A continuación se presentarán los conceptos necesarios para comprender la presente investigación.

2.1.1. Ciencia de datos

La ciencia de datos es un campo multidisciplinario, es decir, es parte de una intersección de varias disciplinas incluyendo fundamentalmente la estadística, matemáticas y ciencias de la computación. En la tabla 2.1 se muestra la relación de las principales disciplinas involucradas en la ciencia de datos.

Disciplina	Descripción
Estadística y Matemáticas	Dentro de esta disciplina, resaltan las herramientas para comprender las tendencias y patrones dentro de los datos, desde métodos descriptivos básicos, hasta modelos predictivos complejos y análisis inferencial.
Computación y Algoritmos	Debido a las limitaciones humanas para procesar grandes cantidades de datos, el uso de hardware y software adecuados se vuelve crucial para poder aumentar la eficiencia, manejar datos complejos y desarrollar algoritmos especializados para el manejo de datos como *Machine Learning* y otras técnicas para la minería de datos.

Tabla 2.1. Disciplinas involucradas en la ciencia de datos

Evolución histórica

La ciencia de datos tiene sus raíces en la estadística y el análisis de datos. Antes de la era de las computadoras, los estadísticos utilizaban métodos manuales para analizar datos, enfocándose principalmente en aplicaciones en las ciencias sociales, biología y economía. Durante este periodo, las investigaciones estaban limitadas a conjunto de datos relativamente pequeños, los cuales eran revisados meticulosamente y se necesitaban de largas horas de cálculos y de análisis para poder encontrar hallazgos significativos y desarrollar modelos que pudieran explicar su comportamiento. Este enfoque manual, aunque detallado, estaba restringido por las limitaciones de tiempo y recursos, lo que a menudo limitaba su alcance y la escala de los análisis estadísticos. fueron estas limitaciones sumado a la proliferación de datos lo que puso en manifiesto la necesidad de desarrollar técnicas de análisis de datos a gran escala (Emmert-Streib et al., 2016).

Con el advenimiento de las computadoras y las mejoras en el almacenamiento de datos, se facilitó la manipulación y análisis de grandes conjuntos de datos. Durante este tiempo, se desarrollaron lenguajes de programación que se convertirían en herramientas fundamentales en la ciencia de datos.

La popularización del internet llevó una explosión en la generación de datos. Las empresas comenzaron a recopilar grandes cantidades de información sobre el comportamiento de clientes, usuarios, transacciones y demás. Este crecimiento exponencial en la generación de datos de manera masiva impulsó la necesidad de métodos más sofisticados para analizar de forma más eficiente y eficaz estos datos.

A principios del siglo XXI, el término Big Data se popularizó para describir conjuntos de datos grandes y complejos, lo cual fue el resultado de el rápido crecimiento tecnológico, la reducción de costes en la fabricación de los dispositivos de almacenamiento, sacando a evidencia nuevos obstáculos con respecto al procesamiento de datos masivos. Era evidente que los métodos tradicionales de procesamiento de datos eran inadecuados para abordar esta nueva problematica. Paralelamente a esto, el Machine Learning comenzó a utilizarse para crear modelos predictivos a partir de estos grandes conjuntos de datos.

Actualmente la ciencia de datos se ha fusionado con la inteligencia artificial, utilizando

técnicas de aprendizaje profundo para analizar datos. Esto ha llevado avances significativos en campos como la medicina personalizada, la optimización de procesos industriales y la inteligencia empresarial.

La aplicación de la ciencia de datos es muy vasta y se encuentran en prácticamente todos los sectores. Desde hasta mejorar la atención médica dentro del análisis predictivo hasta impulsar estrategias comerciales en el marketing digital. La ciencia de datos actualmente está tomando un papel crucial en la resolución de problemas sociales complejos, como el análisis de cambio climático y la gestión de recursos.

2.1.2. Visualización de datos

La generación masiva de datos que alcanza petabytes diarios, el simple procesamiento numérico resulta insuficiente. El cerebro humano es notablemente más eficiente para identificar patrones y tendencias cuando la información se presenta de forma gráfica (El-Din et al., 2020). Por esta razón, la visualización de datos se ha convertido en una herramienta esencial para comprender y analizar grandes volúmenes de información, facilitando la toma de decisiones informadas en diversos ámbitos.

Un ejemplo clásico que ilustra la importancia de la visualización es el 'cuarteto de Anscombe' (1973). Este caso presenta cuatro conjuntos de datos con estadísticas resumidas idénticas que, al graficarse, revelan patrones completamente distintos. Así, se demuestra que confiar únicamente en promedios y varianzas puede conducir a interpretaciones erróneas si no se visualizan adecuadamente los datos (Anscombe, 1973).

Contexto histórico

Aunque a menudo se piensa que la visualización de información cuantitativa es un desarrollo moderno, en realidad tiene raíces históricas profundas. Desde los primeros mapas y diagramas temáticos en siglos pasados, la representación gráfica de datos ha formado parte de la estadística y la ciencia mucho antes de la era digital. Por ejemplo, en el siglo XIX, pioneros como Florence Nightingale emplearon gráficos manuales para obtener hallazgos valiosos. Sus famosos diagramas de área polar de 1857, también

conocidos como 'gráficos de rosa', mostraron de forma contundente que más soldados morían por enfermedades y heridas que en combate durante la Guerra de Crimea.

De manera similar, el médico John Snow trazó en 1854 un mapa de Londres marcando los casos de cólera, lo que le permitió identificar visualmente que los fallecimientos se agrupaban en torno a una bomba de agua contaminada. Este hallazgo señaló el origen del brote y sentó las bases de la epidemiología moderna basada en mapas.

A lo largo de los siglos XIX y XX, los avances en tecnologías de impresión, en técnicas estadísticas y posteriormente en computación, ampliaron el alcance y la complejidad de las visualizaciones disponibles. La llegada de las computadoras en el siglo XX impulsó una verdadera revolución en este campo, ya que surgieron herramientas digitales interactivas que permiten generar gráficos dinámicos y explorar datos de manera inmediata. En décadas recientes, esto dio lugar a la analítica visual (visual analytics), una disciplina que combina métodos analíticos automatizados con interfaces visuales interactivas para aprovechar simultáneamente la capacidad computacional y el razonamiento humano en la exploración de grandes volúmenes de datos (Keim et al., 2008).

Gracias a estos desarrollos históricos, hoy la visualización de datos es un pilar fundamental en el análisis de información, integrando lo mejor de la intuición humana y el poder de cómputo moderno.

2.1.2.1. Técnicas de visualización

En el análisis exploratorio de datos (EDA, por sus siglas en inglés), se emplea un conjunto de técnicas gráficas bien establecidas para resumir y examinar distribuciones, tendencias y relaciones de manera visual. Estas técnicas permiten detectar patrones, valores atípicos (outliers) y estructuras subyacentes que podrían pasar desapercibidas en tablas numéricas (Abbasov, 2023). Entre las principales técnicas se incluyen:

Histogramas. *Gráficos de barras que muestran la distribución de una variable numérica al agrupar los datos en rangos o intervalos. Los histogramas permiten apreciar de un vistazo la forma de la distribución, ya sea simétrica, sesgada, unimodal o multimodal, así*

como la frecuencia de los valores en cada intervalo. Esto facilita la identificación de sesgos o colas largas. Por ejemplo, un histograma puede revelar si los datos se concentran en un rango específico o si existen valores extremos aislados (Abbasov, 2023).

Diagramas de dispersión (*scatter plots*). *Representan pares de datos (x, y) como puntos en un plano cartesiano. Son útiles para explorar relaciones o correlaciones entre dos variables cuantitativas. Al visualizar cada observación como un punto, es posible discernir patrones, como tendencias lineales o no lineales, agrupamientos y la presencia de valores atípicos. Estos diagramas permiten observar si existe correlación positiva, negativa o nula entre dos medidas, complementando la información que aporta el coeficiente de correlación numérico. Es el caso ejemplificado por Anscombe, donde solo al graficar se evidencian diferencias entre conjuntos con igual estadística resumida (Abbasov, 2023).*

Diagramas de caja y bigotes (*boxplots*). *Resumen la distribución de un conjunto de datos numéricos mediante cinco valores clave: mínimo, cuartil inferior, mediana, cuartil superior y máximo. Propuestos por John Tukey en 1977, los boxplots son eficaces para visualizar la dispersión y simetría de los datos, así como para detectar valores atípicos de forma inmediata. La caja central abarca el $50\,\%$ de los datos (del primer al tercer cuartil) con una línea que indica la mediana, mientras que los bigotes se extienden hasta los valores mínimo y máximo no atípicos. Los puntos fuera de este rango se señalan de forma individual. Los boxplots facilitan la comparación de distribuciones entre distintos grupos, resaltando diferencias en la mediana o en la dispersión (Wilcox, 2009).*

Mapas de calor (*heatmaps*). *Utilizan codificación de color en una matriz para representar valores, siendo útiles para visualizar patrones en datos de alta dimensión o en tablas de doble entrada. Cada celda se colorea según la magnitud del valor, siguiendo una escala cromática. Esta técnica es frecuente para examinar matrices de correlación o para datos geoespaciales y genómicos, permitiendo identificar regiones o clústeres con valores similares. Los mapas de calor ayudan a detectar rápidamente tendencias y valores inusuales en grandes volúmenes de datos, apoyando la identificación de relaciones complejas que*

serían difíciles de apreciar numéricamente (Abbasov, 2023).

En conjunto, estas técnicas de visualización constituyen la base del análisis exploratorio de datos. Cada una aporta una perspectiva distinta y, al combinarlas, permiten al analista obtener una comprensión integral del conjunto de datos. Graficar distribuciones y relaciones facilita la formulación de hipótesis informadas, la identificación de problemas de calidad de datos como valores atípicos o sesgos, y la orientación en la selección de modelos o pruebas estadísticas para etapas posteriores del análisis. La literatura reciente señala que el análisis visual exploratorio involucra activamente al analista en el proceso, aprovechando su percepción para descubrir patrones que podrían pasar desapercibidos para los métodos automatizados.

2.1.3. *Aprendizaje automático (Machine Learning)*

El aprendizaje automático se entiende como el conjunto de métodos que permiten a un sistema mejorar su desempeño con la experiencia contenida en los datos. Su auge reciente obedece a la confluencia de avances algorítmicos y teóricos, la disponibilidad de grandes volúmenes de información y el abaratamiento del cómputo. En una revisión panorámica, (Jordan y Mitchell, 2015) ubican al aprendizaje automático en el núcleo de la inteligencia artificial y de la ciencia de datos, con aplicaciones que van desde las ciencias naturales hasta la industria. Por su parte, (Domingos, 2012) sintetiza principios prácticos (como la descomposición representación–evaluación–optimización, la tensión sesgo–varianza y los riesgos de sobreajuste) que orientan el diseño y la validación de modelos. En este trabajo, el aprendizaje automático se presenta como el marco unificador que articula los algoritmos supervisados considerados a continuación, todos con relevancia directa para la exploración rigurosa de los conjuntos de datos. Complementariamente, conviene recordar el trinomio formativo del área (aprendizaje supervisado, no supervisado y por refuerzo), pues en la práctica coexisten en flujos analíticos donde la clasificación y regresión (supervisado) se apoyan frecuentemente en tareas de descubrimiento de estructura (no supervisado) y, en escenarios interactivos, en agentes que optimizan decisiones mediante recompensas acumuladas (aprendizaje por refuerzo).

Árbol de decisión (*Decision Tree*)

Los árboles de decisión inducen recursivamente particiones del espacio de atributos mediante reglas simples (umbrales numéricos o igualdad de categorías) hasta alcanzar hojas con clases o valores relativamente puros. La formulación clásica de (Quinlan, 1986) introduce criterios de selección basados en entropía y ganancia de información, detalla el sistema ID3 y discute mecanismos para lidiar con ruido y datos incompletos. Entre sus ventajas destacan la interpretabilidad y la capacidad de manejar variables mixtas sin requerir estandarización previa. Sus limitaciones son la inestabilidad ante pequeñas perturbaciones de los datos y la tendencia al sobreajuste; no obstante, su eficacia como aprendices base se potencia cuando se ensamblan en métodos de bagging y boosting, así como en bosques aleatorios (Breiman, 2001). En la práctica, la lectura de un árbol equivale a seguir un camino de condiciones desde la raíz hasta una hoja (lo que facilita la trazabilidad de cada predicción y favorece soluciones parsimoniosas cuando los árboles son pequeños); a la vez, el crecimiento sin control suele fragmentar excesivamente los datos y capturar ruido, motivo por el que la poda (pre o post) y la validación cruzada se vuelven mecanismos indispensables para equilibrar sesgo y varianza. Además, resulta útil distinguir las familias históricas: ID3/C4.5 (con umbrales informativos) y CART (con impureza Gini para clasificación y divisiones para regresión); ambas siguen un esquema voraz y, aunque el problema global de aprender el árbol óptimo es NP–completo, los heurísticos alcanzan buen desempeño y constituyen la base de implementaciones modernas.

Regresión logística (*Logistic Regression*)

La regresión logística modela la probabilidad de un resultado binario mediante la función logística aplicada a una combinación lineal de predictores, permitiendo estimación, prueba de hipótesis e interpretación en razones de momios (log-odds). El tratamiento seminal de (Cox, 1958) formaliza el análisis de secuencias binarias y asienta los procedimientos inferenciales modernos. En contextos de muestreo retrospectivo, (Prentice y Pyke, 1979) demuestran la validez de la estimación por máxima verosimilitud con di-

seños de casos y controles, lo que explica su ubicidad en epidemiología. El contraste de bondad de ajuste de (Hosmer y Lemeshow, 1980) se ha convertido en una herramienta estándar para evaluar especificación de modelos. En la práctica, el ajuste se resuelve como optimización convexa y la predicción requiere apenas un producto punto con los coeficientes estimados; la interpretación en odds ratios conserva intuición causal en estudios observacionales (con las debidas cautelas). Cuando hay muchas variables o clases desbalanceadas, la regularización (Ridge o Lasso) y la calibración de probabilidades contribuyen a evitar sobreadaptación y a estabilizar la inferencia.

K vecinos más cercanos (*K-Nearest Neighbors*, KNN)

KNN asigna a una observación el rótulo (o valor) derivado de los K ejemplos más próximos bajo una métrica definida. El resultado clásico de (Cover y Hart, 1967) acota el error de KNN respecto al clasificador de Bayes, justificando su atractivo como método no paramétrico competitivo cuando hay suficientes datos y una métrica adecuada. Sus raíces de consistencia asintótica se remontan al informe técnico de (Fix y Hodges, 1951). En su implementación básica, KNN es aprendizaje basado en instancias (o perezoso): no induce un modelo global, sino que difiere el costo al momento de predecir (cálculo de distancias contra el conjunto de entrenamiento). Esto lo hace sensible a escalas y a atributos irrelevantes, por lo que es habitual estandarizar, seleccionar características o aplicar reducción de dimensionalidad (por ejemplo, antes de comparar en espacios grandes). Asimismo, la complejidad crece con el número de observaciones y la dimensión, y la elección de K y la métrica introduce un compromiso sesgo–varianza que se calibra por validación cruzada.

Máquinas de vectores de soporte (*Support Vector Machines*, SVM)

Las SVM buscan un hiperplano con margen máximo que separe las clases en un espacio de características; mediante funciones kernel es posible construir separaciones no lineales en espacios de dimensión alta manteniendo una formulación convexa. (Cortes y Vapnik, 1995) establecen la formulación canónica con variables de holgura para datos

no separables, y el trabajo de (Boser et al., 1992) introduce el entrenamiento por maximización de margen. La teoría de aprendizaje estadístico ofrece cotas de generalización y justifica el papel central del margen (Vapnik, 1998). En términos prácticos, SVM destaca cuando hay muchas variables y pocos ejemplos, siempre que se elija un kernel adecuado (lineal, polinomial, RBF o sigmoide) y se regule correctamente C (además de parámetros del kernel). Sus desventajas típicas son el costo de entrenamiento en conjuntos masivos, la elección de hiperparámetros y la interpretabilidad limitada frente a modelos lineales o basados en reglas.

Redes neuronales (*Neural Network*)

La popularización moderna de las redes neuronales se debió al redescubrimiento y sistematización de la retropropagación, que permite ajustar pesos por gradiente en arquitecturas multicapa. (Rumelhart et al., 1986) muestran que se pueden aprender representaciones internas útiles para reconocimiento y predicción, y la revisión de (LeCun et al., 2015) sintetiza cómo el aprendizaje profundo (apoyado en grandes datos, aceleración por GPU y regularización) alcanza el estado del arte en visión, voz y otros dominios. En el plano teórico, el resultado de aproximación universal de (Cybenko, 1989) establece que redes de una capa oculta con activaciones sigmoides pueden aproximar funciones continuas bajo supuestos suaves. Históricamente, el campo se remonta a McCulloch y Pitts (neurona formal), Rosenblatt (perceptrón), el impasse señalado por Minsky y Papert (limitaciones de separabilidad lineal) y el posterior renacimiento con la retropropagación (Werbos, y luego Rumelhart–Hinton–Williams). Desde la década de 2010, arquitecturas profundas especializadas (convolucionales, recurrentes) y la disponibilidad de datos y cómputo han impulsado avances notables en clasificación de imágenes, traducción automática y reconocimiento de voz; a cambio, crecen los retos de consumo de datos y de interpretabilidad.

Bosque aleatorio (*Random Forest*)

Los bosques aleatorios combinan múltiples árboles entrenados sobre muestras bootstrap e introducen aleatoriedad en la selección de variables candidatas por división, lo que reduce la correlación entre árboles y, por ende, la varianza del ensamble. (Breiman, 2001) demuestran la convergencia del error de generalización y popularizan estimadores internos como el error out-of-bag (OOB) y las medidas de importancia de variables. En la práctica, ofrecen buen desempeño con bajo ajuste de hiperparámetros y robustez frente a variables irrelevantes; su principal limitación es la interpretabilidad a nivel global, mitigable con herramientas de explicación. De manera operativa, la combinación de bagging y el método de subespacio aleatorio (selección de un subconjunto de atributos por nodo) promueve bosques de árboles poco correlacionados cuyo promedio estabiliza la predicción; el error OOB proporciona una estimación interna de desempeño y las importancias por permutación aportan señales útiles sobre relevancia de atributos para auditoría y toma de decisiones.

Árboles potenciados (*Boosted Trees* / *Gradient Boosting*)

El boosting agrega secuencialmente aprendices débiles para corregir errores residuales de etapas previas, construyendo un modelo aditivo que minimiza una función de pérdida. (Friedman, 2001) formaliza el gradient boosting como una aproximación voraz en el espacio de funciones y muestra su eficacia cuando el aprendiz base es un árbol de decisión; más tarde, (Friedman, 2002) introduce la variante estocástica, que mejora robustez y eficiencia al muestrear datos por iteración. En paralelo, AdaBoost (Freund y Schapire, 1997) proporciona una perspectiva de reponderación con garantías teóricas. Su rendimiento depende de una regularización cuidadosa (tasa de aprendizaje, profundidad y número de iteraciones), la vigilancia del sobreajuste mediante validación y el uso de criterios de parada temprana.

Ensamble heterogéneo (*Stacking*/*Blending*)

El stacking integra modelos de distinta naturaleza a través de un metaaprendiz que aprende, con validación cruzada, a combinar las predicciones base para minimizar el error de generalización. (Wolpert, 1992) introduce el esquema y clarifica el rol del segundo nivel entrenado sobre predicciones de validación, mientras que Super Learner (van der Laan et al., 2007) formaliza una versión con garantías tipo oráculo bajo pérdidas generales. Su potencia radica en explotar sesgos y varianzas complementarias; sus riesgos, las fugas de información si no se respeta el apilamiento con validación cruzada y la complejidad operativa del despliegue.

2.1.4. Inteligencia artificial generativa (IA generativa)

La inteligencia artificial generativa se enfoca en la creación de contenido nuevo (texto, imágenes, audio, etc.) a partir de modelos de IA entrenados en grandes conjuntos de datos. A diferencia de los enfoques discriminativos (que reconocen o clasifican patrones en datos existentes), los modelos generativos aprenden la distribución estadística de los datos de entrenamiento para poder generar nuevas muestras que resulten verosímiles respecto a los datos originales. Este paradigma ha cobrado gran relevancia recientemente gracias a los avances en algoritmos y potencia de cómputo, permitiendo aplicaciones como la síntesis de imágenes realistas, la composición musical automática y los modelos de lenguaje que producen texto coherente.

Historia y antecedentes de la IA generativa

Los primeros esfuerzos en IA generativa incluyen modelos probabilísticos clásicos en aprendizaje automático, pero los mayores hitos llegaron con la era del deep learning. En la década de 2010 surgieron modelos generativos profundos que demostraron avances significativos. Por ejemplo, los autoencoders variacionales y las redes generativas antagónicas (conocidas como generative adversarial networks, o GANs) permitieron por primera vez la generación de imágenes sintéticas de apariencia realista. En particular,

la introducción de las GAN por Goodfellow et al. en 2014 mostró cómo un modelo generativo puede aprender a producir datos nuevos enfrentando a dos redes neuronales en un juego minimax (generador vs. discriminador) (Goodfellow et al., 2014). Estos antecedentes sentaron las bases para la explosión reciente de la IA generativa en múltiples dominios.

El *Transformer* y el artículo *Attention Is All You Need*

Un punto de inflexión en el campo de la IA generativa, particularmente en el procesamiento de lenguaje natural, fue la introducción de la arquitectura Transformer en 2017 (Vaswani et al., 2017). El influyente artículo "Attention Is All You Need" (Vaswani et al., 2017) presentó un modelo de red neuronal basado exclusivamente en mecanismos de atención ("self-attention"), eliminando la necesidad de recurrencia o convoluciones en la construcción de secuencias. Este diseño permitió entrenar modelos de secuencia de manera más paralela y eficiente, capturando dependencias de largo alcance en el texto con mayor eficacia. Los Transformers lograron resultados superiores en tareas de traducción automática y otras tareas de secuencias, estableciendo una nueva arquitectura estándar para modelos generativos de lenguaje. Desde entonces, el Transformer se ha convertido en la piedra angular de los modelos de IA generativa modernos, sirviendo de base tanto para modelos de lenguaje como para modelos generativos de imágenes (p.,ej., en visión por computadora).

Modelos de lenguaje grandes (LLMs)

Siguiendo el éxito del Transformer, la investigación en generación de lenguaje se orientó a escalar el tamaño de los modelos y la cantidad de datos de entrenamiento. Así nacieron los llamados modelos de lenguaje grandes (Large Language Models, LLMs), que son redes neuronales con miles de millones de parámetros entrenadas sobre corpora masivos de texto (Zhao et al., 2023). Al aumentar la escala, estos modelos demostraron mejoras sustanciales en desempeño e incluso comportamientos emergentes que no se veían en modelos más pequeños. Un LLM típico emplea la arquitectura

Transformer preentrenada para predecir la siguiente palabra en una secuencia, habiendo absorbido durante su entrenamiento conocimiento amplio de lenguaje. Gracias a su gran capacidad, los LLMs pueden generar textos coherentes y contextualmente relevantes, traducir entre idiomas, responder preguntas y realizar una variedad de tareas sin ajustes específicos. Un ejemplo destacado es ChatGPT de OpenAI, un sistema conversacional basado en un LLM que ha logrado mantener diálogos naturales y resolver instrucciones complejas, lo que demuestra el potencial de estas tecnologías y ha popularizado la IA generativa a nivel masivo (su lanzamiento en 2022 atrajo la atención mundial hacia los LLM (Zhao et al., 2023)).

OpenAI GPT: evolución y conceptos

La serie GPT de OpenAI (acrónimo de Generative Pre-trained Transformer, es decir, Transformador Generativo Preentrenado) ejemplifica la evolución reciente de los modelos generativos de lenguaje. GPT-1, introducido en 2018, demostró que preentrenar un modelo Transformer con una gran cantidad de texto y luego ajustarlo finamente a tareas específicas era una estrategia muy efectiva para el aprendizaje no supervisado de representaciones lingüísticas. GPT-2 (2019) amplió drásticamente el tamaño del modelo y mostró una capacidad sorprendente para generar texto coherente y continuo, hasta el punto de plantear preocupaciones sobre la diseminación de desinformación generada por máquinas. Posteriormente, GPT-3 (2020) elevó el número de parámetros a 175 mil millones, lo que habilitó el fenómeno del aprendizaje con pocas muestras (few-shot learning): GPT-3 puede adaptarse a múltiples tareas simplemente a partir de unos pocos ejemplos o instrucciones en texto, sin entrenamiento adicional específico (Brown et al., 2020). Más recientemente, GPT-4 (2023) continuó la tendencia de incremento de escala e introdujo capacidades multimodales (aceptando entrada de texto e imagen), logrando un rendimiento aún más avanzado en comprensión y generación de lenguaje (OpenAI, 2023). La evolución de GPT refleja varios conceptos clave de la IA generativa moderna: el aprovechamiento del preentrenamiento masivo, el uso de arquitecturas basadas en atención, el escalado de modelos para obtener nuevas habilidades emergentes y la

incorporación de técnicas como el ajuste fino con retroalimentación humana (RLHF) para alinear la generación de texto con las intenciones del usuario. En conjunto, los modelos GPT han impulsado gran parte del progreso en aplicaciones de lenguaje natural, desde asistentes virtuales y chatbots hasta sistemas de resumen automático y generación de código fuente, consolidando el papel de la IA generativa en la actualidad.

2.2. Literatura relacionada

Tang, Yang, Fan, Cao, Luo, y Halevy (2023), en su artículo "VerifAI: Verified Generative AI", publicado en arXiv, discuten las preocupaciones crecientes sobre la precisión y fiabilidad de los resultados de la IA generativa. autores proponen VerifAI, un sistema para verificar los resultados de la IA generativa desde una perspectiva de gestión de datos. Este enfoque implica analizar los datos subyacentes de lagos de datos multimodales, como archivos de texto, tablas y grafos de conocimiento, para evaluar su calidad y consistencia. El objetivo es establecer una base más sólida para evaluar los resultados de los modelos de IA generativa, asegurando su corrección, promoviendo la transparencia y permitiendo la toma de decisiones con mayor confianza. El artículo destaca la importancia de verificar la IA generativa para contribuir a un uso más confiable y responsable de la IA (Tang et al., 2023).

En un artículo titulado "Can artificial intelligence help for scientific writing?"publicado en Critical Care (2023), los autores Michele Salvagno, Fabio Silvio Taccone y Alberto Giovanni Gerli exploran el uso de ChatGPT, un LLM desarrollado por OpenAI, en la escritura científica. El artículo discute cómo ChatGPT y otras IA generativas de tipo LLM pueden ser herramientas útiles en la organización de material, generación de borradores iniciales y revisión de textos en el ámbito de la escritura científica. Sin embargo, enfatizan que estos no deben reemplazar el juicio humano y siempre deben ser revisados por expertos antes de su uso en decisiones críticas.

Los autores también abordan preocupaciones éticas sobre el uso de estas herramientas, como el riesgo de plagio, inexactitudes y posibles desequilibrios en su accesibilidad entre países de ingresos altos y bajos. Concluyen que se requerirá un consenso sobre

cómo regular el uso de IA generativa en la escritura científica y establecer mecanismos para identificar y penalizar el uso no ético. Además, sugieren que ChatGPT podría tener aplicaciones prácticas en la práctica clínica como herramienta de ahorro de tiempo, aunque subrayan la necesidad de la supervisión humana en el proceso de escritura científica (Salvagno et al., 2023).

En el artículo "Demonstration of InsightPilot: An LLM-Empowered Automated Data Exploration System"de Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, y Dongmei Zhang, publicado en arXiv en 2023, se presenta InsightPilot, un sistema automatizado de exploración de datos impulsado por un LLM. El sistema está diseñado para ayudar a los usuarios a comprender e interpretar conjuntos de datos de manera más efectiva, seleccionando automáticamente las intenciones"de análisis adecuadas, como comprender, resumir y explicar. Estas intenciones de análisis se concretizan mediante consultas intencionales llamadas IQueries para crear una secuencia de exploración coherente y significativa. InsightPilot emplea el modelo GPT-3.5-Turbo. El artículo demuestra la efectividad de InsightPilot en un estudio de caso, mostrando cómo puede ayudar a los usuarios a obtener hallasgos significativos de sus conjuntos de datos. Este trabajo menciona de igual forma limitaciones existentes que hay con respecto a las herramientas de exploración de datos al no ser capaces de comprender la intención del usuario o el conntexto del conjunto de datos. Propone superar estas limitaciones al proporcionar una herramienta automatizada que mejora la eficiencia y que de igual forma reduzca el esfuerzo manual en la exploración de datos (Ma et al., 2023).

En el artículo "FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis"de Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon y Xiaofeng Zhang, publicado en 2023, se presenta FinVis-GPT, un modelo de lenguaje grande multimodal (LLM) diseñado específicamente para el análisis de gráficos financieros. FinVis-GPT aprovecha el poder de los LLM y combina la sintonización de instrucciones y capacidades multimodales para interpretar gráficos financieros y proporcionar análisis valiosos. Para entrenar FinVis-GPT, se generó un conjunto de datos orientado a tareas financieras, que incluye varios tipos de gráficos financieros y sus descripciones correspondientes. El rendimiento del modelo se evaluó a través de varios estudios de caso, y los resultados prometedores

demostraron que FinVis-GPT supera a los modelos multimodales LLM existentes en tareas relacionadas con gráficos financieros. FinVis-GPT representa un esfuerzo pionero en el uso de LLM multimodales en el dominio financiero (Wang et al., 2023).

En el artículo "Using GPT-4 to Augment Unbalanced Data for Automatic Scoring"de Luyang Fang, Gyeong-Geon Lee, Xiaoming Zhai y colaboradores, publicado en 2023, introduce un marco de trabajo novedoso para la ampliación de datos de texto utilizando GPT-4. Este estudio resalta el potencial y la efectividad de las técnicas de ampliación de datos utilizando modelos de lenguaje generativos de gran tamaño, como GPT-4, para abordar conjuntos de datos desequilibrados en la evaluación automatizada (Fang et al., 2023).

En el artículo "ChatGPT as your Personal Data Scientist"de Md Mahadi Hassan, Alex Knipper y Shubhra Kanti Karmaker Santu, publicado en 2023, se explora el uso de ChatGPT como una herramienta para facilitar tareas de aprendizaje automático (AutoML) a través de conversaciones naturales e intuitivas. El artículo aborda el desafío de comprender datos específicos del dominio y definir tareas de predicción, lo que generalmente requiere intervención humana y hace que el proceso sea tedioso y no completamente automatizado. Los autores proponen un marco de ciencia de datos conversacional basado en ChatGPT que actúa como un çientífico de datos personal", utilizando Modelos de Lenguaje Grande (LLM) para construir una interfaz natural entre los usuarios y los modelos de aprendizaje automático (como Scikit-Learn). El modelo se centra en cuatro estados de diálogo: Visualización de Datos, Formulación de Tareas, Ingeniería de Predicción y Resumen de Resultados y Recomendaciones. Cada estado representa una fase única en la conversación, impactando la interacción general usuario-sistema. El sistema desarrollado no solo demuestra la viabilidad del concepto de ciencia de datos conversacional, sino que también subraya la potencia de los LLM en la resolución de tareas complejas. Además, el desarrollo del sistema reveló varias debilidades críticas en los LLM actuales (ChatGPT) y destacó oportunidades sustanciales para mejoras (Hassan et al., 2023).

2.3. Marco tecnológico

Se detallan las herramientas tecnológicas utilizadas en la investigación, así como el entorno de trabajo y las tecnologías empleadas para el análisis de datos. Estas herramientas facilitan la implementación de las etapas del ciclo de vida CRISP-DM, tanto mediante métodos tradicionales como mediante la IA generativa.

2.3.1. Herramientas de Programación y Análisis de Datos

2.3.1.1. Lenguaje R

El lenguaje R es ampliamente utilizado en la ciencia de datos y la estadística debido a su capacidad para manejar y analizar grandes conjuntos de datos de manera eficiente. Es conocido por su gran cantidad de paquetes y librerías, que proporcionan herramientas especializadas para diversas tareas de análisis de datos.

Paquetes Utilizados:

- ***tidyverse**: Es una colección de paquetes R diseñada para la ciencia de datos. Incluye ggplot2 para la visualización de datos, dplyr para la manipulación de datos, tidyr para la limpieza de datos, readr para la importación de datos, entre otros.*
- ***ggplot2**: Utilizado para crear visualizaciones elegantes y complejas de datos.*
- ***dplyr**: Facilita la manipulación de datos, permitiendo realizar operaciones como filtrado, selección, agrupamiento y resumen de datos.*
- ***tidyr**: Ayuda a limpiar y dar forma a los datos para que sean más fáciles de analizar.*
- ***caret**: (Classification And Regression Training) Este paquete simplifica el proceso de creación de modelos de machine learning, proporcionando funciones para el preprocesamiento de datos, la selección de modelos y la evaluación del rendimiento.*
- ***randomForest**: Utilizado para la clasificación y regresión mediante la construcción de múltiples árboles de decisión.*

- ***e1071****: Proporciona funciones para máquinas de vectores de soporte (SVM) y otros métodos estadísticos.*
- ***nnet:*** *Implementa redes neuronales de retropropagación y modelos de regresión logística.*

2.3.1.2. Jupyter Notebooks

Jupyter Notebooks es una herramienta de código abierto que permite crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto narrativo. Es ampliamente utilizado en la investigación y educación en ciencia de datos debido a su flexibilidad y facilidad de uso.

Características Principales*:*

- ***Integración Multilenguaje****: Permite ejecutar código en múltiples lenguajes de programación, incluyendo Python, R y Julia.*
- ***Documentación Interactiva****: Facilita la combinación de código ejecutable con texto explicativo, ecuaciones matemáticas y visualizaciones, creando un entorno de trabajo interactivo.*
- ***Modularidad****: Permite dividir el análisis en celdas independientes, cada una de las cuales puede contener código, texto, o visualizaciones. Esto facilita la organización y el seguimiento del proceso de análisis.*
- ***Visualización de Datos****: Integra bibliotecas de visualización como matplotlib, seaborn, plotly, entre otras, permitiendo crear gráficos y visualizaciones interactivas.*

2.3.1.3. Anaconda

Anaconda es una distribución de Python y R diseñada para la ciencia de datos y el aprendizaje automático. Incluye más de 1,500 paquetes seleccionados y la herramienta de gestión de entornos Conda, que facilita la instalación y gestión de dependencias.

Ventajas de Uso*:*

- ***Gestión de Entornos Virtuales**: Permite crear y gestionar entornos virtuales, asegurando que cada proyecto tenga sus propias versiones de paquetes y librerías, lo que mejora la reproducibilidad y evita conflictos entre dependencias.*
- ***Incluye Herramientas Populares**: Viene con Jupyter Notebooks, RStudio, y otras herramientas esenciales para la ciencia de datos preinstaladas.*
- ***Instalación Simplificada**: Facilita la instalación y actualización de paquetes mediante el gestor de paquetes Conda, lo que reduce las complicaciones asociadas a la configuración del entorno de desarrollo.*

2.3.2. Herramientas de Inteligencia Artificial

API de OpenAI

La API de OpenAI proporciona acceso a modelos de lenguaje de gran escala, como GPT-4o, que son capaces de generar texto, responder preguntas, realizar traducciones y más. Estos modelos son fundamentales para implementar la IA generativa en este estudio.

- ***Procesamiento avanzado de lenguaje natural**. GPT-4o puede comprender instrucciones complejas, mantener contexto a largo plazo y generar texto coherente alineado con distintos estilos (académico, técnico, explicativo). Esto permite automatizar la redacción de análisis, descripciones metodológicas y documentación completa dentro del ciclo CRISP-DM.*
- ***Generación de código en múltiples lenguajes**. Produce código funcional en R, Python, SQL y otros lenguajes comúnmente utilizados en ciencia de datos. Puede generar funciones, pipelines, *scripts* completos, visualizaciones y módulos reutilizables a partir de instrucciones textuales, además de corregir, depurar y optimizar el código generado.*
- ***Lectura, interpretación y transformación de datos**. GPT-4o puede describir estructuras de datos, identificar inconsistencias conceptuales, sugerir transformacio-*

nes adecuadas (normalización, codificación, imputación) y estructurar pasos de preprocesamiento basándose únicamente en la descripción textual del usuario o en fragmentos de datos introducidos en el prompt.

- ***Modelado predictivo automatizado**. El modelo es capaz de generar *scripts* que realizan entrenamiento, ajuste de hiperparámetros, evaluación y comparación de modelos (árboles de decisión, *boosting*, bosques aleatorios, SVM, redes neuronales, entre otros), produciendo salidas reproducibles.*
- ***Multimodalidad nativa**. GPT-4o permite trabajar de manera integrada con texto, imágenes, audio y, en ciertos entornos, video. Puede leer gráficos, tablas, capturas de pantalla, manuscritos o resultados, interpretarlos y transformarlos en código o análisis estructurados. Esto facilita casos prácticos como extraer datos desde imágenes de tablas o interpretar resultados mostrados en pantallas de herramientas externas.*
- ***Razonamiento estructurado y planificación**. Puede construir planes de análisis, descomponer tareas complejas en pasos ejecutables, seleccionar algoritmos adecuados según la naturaleza del problema y explicar las decisiones tomadas. Esta capacidad resulta útil en la automatización de etapas completas del proceso CRISP-DM.*
- ***Generación de artefactos reproducibles**. GPT-4o puede producir documentos completos en formatos como R Markdown, Jupyter Notebooks u otros formatos ejecutables, integrando texto, código, visualizaciones y conclusiones. Esto permite crear reportes reproducibles en una sola interacción.*
- ***Construcción de agentes autónomos**. El modelo puede emplearse como núcleo de agentes conversacionales que realizan secuencias de acciones, consultan al usuario, generan código, ejecutan análisis y devuelven artefactos completos. Puede coordinar tareas, mantener memoria dentro de la sesión y operar como interfaz de exploración analítica para usuarios no técnicos.*

- ***Comprensión y depuración de errores**. GPT-4o puede leer trazas de errores, diagnosticar causas probables y proponer soluciones específicas en código. Esto es especialmente útil en flujos iterativos de ciencia de datos donde la depuración es frecuente.*
- ***Simulación de escenarios y análisis comparativo**. Puede explicar diferencias de rendimiento entre modelos, interpretar métricas, analizar sensibilidad–especificidad y sugerir mejoras metodológicas, siguiendo criterios comunes en el análisis estadístico y la evaluación de modelos.*

2.3.3. Entorno de Trabajo

Para llevar a cabo los experimentos y comparaciones, se utilizaron entornos virtuales específicos configurados mediante Anaconda. Estos entornos aseguran la reproducibilidad de los resultados y la gestión eficiente de las dependencias de software.

***Configuración del Entorno**:*

- ***Creación de Entornos Virtuales**: Se crearon entornos virtuales separados para la ejecución de scripts en R y para el uso de herramientas de IA generativa. Esto garantiza que cada enfoque tenga un entorno optimizado y libre de conflictos de dependencias.*
- ***Instalación de Paquetes**: Se instalaron las versiones específicas de los paquetes y librerías necesarias para el análisis y modelado de datos, asegurando compatibilidad y estabilidad en el proceso de investigación.*
- ***Documentación en Jupyter Notebooks**: Todo el proceso de análisis y los resultados fueron documentados en Jupyter Notebooks, proporcionando un registro detallado y accesible del flujo de trabajo.*

Capítulo 3

Implementación del ciclo CRISP-DM: Humano vs IA

3.1. Metodología general de la experimentación

3.1.1. Descripción de los Datasets

En este apartado se describen los seis conjuntos de datos utilizados en el estudio, resaltando sus características principales y atributos clave. Todos los conjuntos de datos seleccionados provienen del área de la salud y fueron elegidos por su relevancia y diversidad en cuanto a tamaño, número de variables, presencia de valores faltantes y balance de clases. Esta diversidad garantiza una evaluación robusta de las metodologías de exploración de datos, al abarcar distintos escenarios y retos asociados a los datos.

Autistic Spectrum Disorder Screening Data for Children

Este conjunto de datos corresponde a la detección del Trastorno del Espectro Autista (TEA; en inglés Autism Spectrum Disorder, ASD) en población infantil, obtenidos a través del cuestionario de cribado AQ-10-Child. El conjunto de datos tiene una dimensionalidad de 292 observaciones por 21 variables, los cuales incluyen indicadores conductuales y variables demográficas. Una característica notable es la presencia de valores faltantes, lo que exige la aplicación de técnicas de imputación o un manejo específico de datos

incompletos durante la etapa de preprocesamiento. La variable objetivo es binaria (positivo/negativo para TEA) y presenta una distribución de clases casi equilibrada, con un 48.29 % de casos positivos frente a un 51.71 % de negativos. Este balance reduce el riesgo de sesgo en el modelo y facilita una evaluación imparcial del rendimiento de los algoritmos de clasificación.(Thabtah, 2017).

Epileptic Seizure Recognition Data Set

Este conjunto de datos corresponde al reconocimiento de episodios de epilepsia a partir de datos extraídos de señales EEG (electroencefalograma). Posee un tamaño considerable, con 11,500 observaciones registradas, cada una descrita mediante 180 variables que representan medidas o transformaciones de las señales cerebrales en diferentes canales y momentos. No existen valores faltantes en este conjunto de datos, lo que significa que todas las observaciones tienen completos sus 180 variables. Sin embargo, la distribución de clases es altamente desequilibrada: alrededor de 80 % de las instancias corresponden a casos negativos (sin convulsión) y solo 20 % a casos positivos (con presencia de una convulsión epiléptica). Esta disparidad 4:1 indica que la clase negativa domina el conjunto de datos, por lo que al aplicar técnicas de modelado y exploración será fundamental abordar este desbalance (por ejemplo, mediante técnicas de muestreo o ajustes en las métricas) para garantizar que las detecciones de convulsiones (clase minoritaria) no pasen desapercibidas.

Diabetes Data Set

Este conjunto de datos, enfocado en la diabetes mellitus, comprende 520 observaciones y 17 variables. Las variables son de naturaleza clínica y abarcan un espectro de mediciones fisiológicas, resultados de pruebas de laboratorio y antecedentes de los pacientes. Un aspecto importante es que no presenta valores faltantes, es decir, todos los registros están completos en cada variable, lo cual simplifica la fase de preparación de datos al no requerir imputación. La distribución de la clase objetivo (presencia de diabetes vs. ausencia) está moderadamente desequilibrada: aproximadamente 38.46 % de las

instancias son negativas (individuos sin diagnóstico de diabetes) y 61.54 % son positivas (casos con diabetes). Esto implica que la clase positiva (diabéticos) es más frecuente, casi 1.6 veces la negativa, por lo que los algoritmos de clasificación deben calibrarse para no inclinarse excesivamente hacia la predicción de la clase mayoritaria y mantener buen rendimiento en la minoritaria.

La selección de estos tres conjuntos de datos proporciona un banco de pruebas diverso para comparar la exploración de datos mediante IA generativa con la de un experto humano. La variabilidad en tamaños (desde cientos hasta miles de observaciones), en número de variables (desde decenas hasta cientos de atributos), en la presencia/ausencia de datos faltantes, y en el equilibrio de las clases, garantiza que cada enfoque sea evaluado bajo distintos escenarios representativos. Esta diversidad permite observar cómo cada método se adapta a desafíos específicos (como manejar alta dimensionalidad, datos incompletos o clases desbalanceadas) y asegura que las conclusiones obtenidas sobre su desempeño sean robustas y generalizables a múltiples contextos en el ámbito de la ciencia de datos.

3.1.2. Selección de los modelos de Aprendizaje Automático

En este estudio comparativo entre una IA generativa y un experto humano (en lenguaje R), se ha optado por incluir una gama amplia de algoritmos de clasificación clásicos. El motivo es cubrir distintos niveles de complejidad e interpretabilidad, de modo que la comparación sea lo más completa posible.

A continuación se justifica la inclusión de cada uno:

Árbol de Decisión

Este algoritmo se seleccionó como un modelo de referencia (baseline) por su alta interpretabilidad. Su principal ventaja radica en la capacidad de descomponer cada predicción en un conjunto de reglas lógicas de tipo si-entonces, las cuales son fácilmente comprensibles para audiencias no especializadas, aportando así una notable transparencia al proceso de toma de decisiones. Gracias a su entrenamiento rápido y bajo costo

computacional, este algoritmo sirve como un modelo de referencia para evaluar la ganancia en precisión de arquitecturas más complejas (Rabanal, 2024). Además, la estructura jerárquica descubre interacciones no lineales entre atributos sin requerir codificación ni imputación previa, preservando así la fidelidad de la muestra. La poda y el control de profundidad mantienen el equilibrio sesgo–varianza antes de escalar a ensambles o modelos más complejos (Alaminos-Fernández, 2023).

Regresión Logística

Este modelo se seleccionó como modelo de referencia, la cual es fundamental para la clasificación binaria por su bajo costo computacional y su capacidad de ofrecer estimaciones probabilísticas fácilmente interpretables que resultan útiles para fijar umbrales clínicos. La regularización $L1$ (penalización absoluta) y $L2$ (penalización cuadrática) controlan el sobreajuste sin sacrificar transparencia, lo que permite comparar de manera rigurosa el desempeño de modelos más complejos frente a una base sólida y explícita (Javier, 2025).

K Nearest Neighbor (KNN)

Este modelo se incluyó como modelo de referencia no paramétrica porque toma decisiones basadas en similitud local, una lógica análoga al razonamiento clínico basado en casos. Al posponer el cómputo hasta la inferencia, acepta nuevos ejemplos sin reentrenar el modelo, rasgo valioso en flujos de datos actualizables. El ajuste de k mediante validación cruzada evidencia de forma transparente el equilibrio sesgo-varianza y establece un modelo de referencia sencillo frente a métodos más sofisticados (IBM, 2025).

Máquina de Vectores de Soporte (SVM)

El modelo SVM fue elegido por su capacidad de maximizar el margen entre clases, lo que favorece la generalización en espacios de alta dimensionalidad, común cuando el número de variables supera al de observaciones. Los kernels permiten incorporar no linealidad sin comprometer fundamentos teóricos, y el hecho de que solo dependa de los

vectores de soporte reduce el uso de memoria. La penalización diferenciada facilita el manejo de clases desbalanceadas, frecuente en los conjuntos biomédicos bajo estudio.

Red Neuronal Artificial

Este tipo de modelo funciona como techo de complejidad y rendimiento dentro del conjunto evaluado. Las redes multicapa capturan interacciones jerárquicas y no lineales que otros modelos omiten, y su escalabilidad con GPU/TPU habilita el procesamiento de grandes volúmenes de datos. El elevado número de hiperparámetros requiere regularización y validación exhaustiva, por lo que sirve para contrastar el impacto de la potencia predictiva frente a la pérdida de interpretabilidad.

Bosque Aleatorio

Este algoritmo fue seleccionado por su notable robustez ante ruido y valores atípicos. Esta característica se deriva de su arquitectura de ensamble, que promedia las predicciones de múltiples árboles de decisión mediante la técnica de bagging. Para garantizar la diversidad entre los árboles, cada uno se entrena sobre una muestra de datos distinta (generada por bootstrap) y, en cada nodo, solo se considera un subconjunto aleatorio de atributos para realizar la división.

Árbol Potenciado

El modelo de Árbol Potenciado (Gradient Boosting) se eligió por su capacidad para alcanzar un rendimiento predictivo superior en problemas con datos tabulares de complejidad intermedia a alta. Su mecanismo de ensamble, basado en una combinación secuencial de árboles débiles, reduce el sesgo al corregir iterativamente los errores residuales. De forma simultánea, mitiga el sobreajuste mediante regularización interna (ej. profundidad limitada, tasa de aprendizaje) y detención temprana. Implementaciones eficientes como XGBoost y LightGBM garantizan su viabilidad en conjuntos de datos a gran escala.

Ensamble heterogéneo

Finalmente, para establecer una cota superior de rendimiento, se diseñó un ensamble heterogéneo. Este meta-modelo integra un conjunto diverso de clasificadores que operan bajo distintos paradigmas algorítmicos, modelos de referencia como Árbol de Decisión, Regresión Logística, K Nearest Neighbor, Máquinas de Vectores de Soporte y Redes Neuronales.

Fundamentado en el principio de diversidad, el ensamble utiliza métodos de votación o apilamiento para combinar las predicciones. Esta estrategia permite que los errores no correlacionados de los modelos base se compensen entre sí, resultando en un predictor final con mayor robustez y capacidad de generalización. Así, el rendimiento de este ensamble no solo sirve como un benchmark práctico, sino que también cuantifica la sinergia obtenida mediante la hibridación de modelos frente al desempeño de cada clasificador de forma individual.

3.2. Exploración de datos (Experto Humano)

Figura 3.1. Etapas del ciclo de ciencia de datos.

La fase de exploración constituye el primer eslabón operativo del ciclo de ciencia de

datos ilustrado en la Figura 3.1, donde el conocimiento preliminar de los datos alimenta la preparación, el modelado y la evaluación, para culminar en la comunicación de resultados y reiniciar iterativamente el proceso. Basados en ese esquema, esta sección describe los análisis exploratorios realizados sobre tres conjuntos de datos descritos en la Sección 3.1.1.

La exploración se ejecutó mediante dos flujos de trabajo paralelos:

- ***Experto humano.*** *Un analista con experiencia en ciencia de datos que condujo un EDA manual, formulando hipótesis preliminares y seleccionando transformaciones basadas en su criterio profesional.*
- ***IA generativa.*** *En el cual se empleó un prompt estandarizado que instruye al modelo a producir un documento en R Markdown conteniendo descripciones, matrices de correlación, perfiles de valores faltantes y visualizaciones clave. Este documento sirve como insumo directo para la etapa de preparación y, posteriormente, para el modelado.*

Ambos enfoques generan reportes comparables, lo que facilita el análisis crítico de convergencias y divergencias antes de las etapas de limpieza, selección de variables y construcción de modelos.

La Figura 3.2 ilustra el proceso seguido por el analista, desde la inspección inicial de la estructura del conjunto hasta la generación de hipótesis que orientan la ingeniería de características.

3.2.1. Autistic Spectrum Disorder Screening Data for Children

El análisis exploratorio de datos realizado por el experto humano sobre el conjunto Autistic Spectrum Disorder Screening Data for Children siguió una metodología estructurada orientada a identificar aspectos clave del conjunto de datos que podrían influir en las etapas posteriores del estudio. A continuación, se describen únicamente los hallazgos derivados de la exploración inicial.

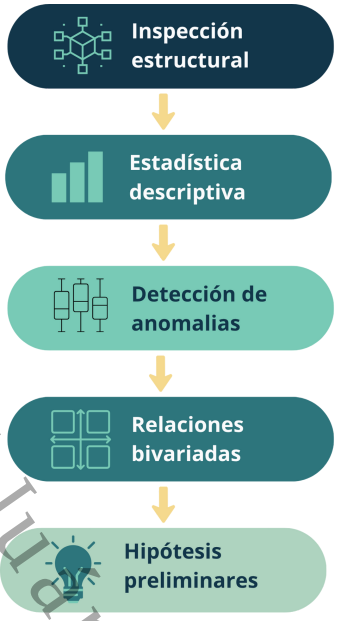

Figura 3.2. Flujo de análisis exploratorio de datos ejecutado por el experto humano.

Inspección estructural

La etapa inicial del estudio consistió en una inspección exploratoria de los datos, enfocada en validar su integridad estructural y caracterizar su contenido. Primeramente, se verificó que el conjunto de datos cargado en memoria correspondiera a las dimensiones proyectadas, confirmando una estructura de 292 observaciones por 15 variables.

Con el objetivo de sistematizar el análisis, las 15 varibales del conjunto de datos se agruparon en tres dominios conceptuales:

a. *Variables de Comportamiento (10 variables): Corresponden a las puntuaciones de comportamiento (AQ-10-Child).*

b. *Variables Demográficas (4 variables):* `age` *(edad del participante),* `gender` *(género),* `jundice` *(presencia de ictericia al nacer) y* `austim` *(antecedente de familiar con autismo).*

c. *Variable Objetivo (1 variable): La variable dependiente clase, que representa la etiqueta de clasificación binaria ('Positivo' o 'Negativo').*

Durante la inspección se confirmó cómo se importó cada columna desde el archivo

CSV: las que contienen letras se leyeron como texto (character) y las numéricas como enteros o números reales (integer/numeric). La columna `clase`*, al traer valores como 'Negativo' y 'Positivo', también se importó como texto; metodológicamente, debe tratarse como una variable categórica binaria. Por ello se dejó asentado (para la fase de preparación de datos) que* `clase`*,* `gender`*,* `jundice` *y* `austim`*, se deben codificar explícitamente como factores con niveles definidos (por ejemplo, Negativo/Positivo; no/yes; femenino/masculino).*

Esta codificación en factor no se debe aplicar aún en la etapa de exploración, pero es necesaria después para que los modelos y resúmenes estadísticos identifiquen correctamente los niveles, mantengan el mismo orden en todo el flujo y eviten inconsistencias entre conjuntos de entrenamiento y prueba.

Antes de proponer cualquier conversión, se verificó la existencia de valores faltantes a nivel global y estratificando por `clase`*. El conteo mostró únicamente 4 ausentes en total, todos en la variable* `age`*, distribuidos de forma simétrica entre Positivo (2 observaciones) y Negativo (2 observaciones) según la Tabla 3.1. Con base en ello, se documentó, que para la siguiente etapa, debe ejecutarse una imputación sencilla y estratificada de* `age` *mediante la mediana dentro de cada clase, y la posterior conversión a factor de las variables categóricas. En esta fase de exploración no se realizaron transformaciones a los datos; únicamente se detectaron y justificaron las acciones a implementar en la fase de preparación de datos.*

Detección de anomalías

Criterio usado. Se aplicó la regla del rango intercuartílico (IQR) para detectar valores atípicos únicamente donde son necesarios: variables numéricas con escala continua u ordinal con suficiente número de valores distintos.

Para cada variable X con observaciones $x_1, \ldots, x_n$ se calculó:

$$\mathrm{IQR} = Q_3 - Q_1, \qquad L_I = Q_1 - 1.5\,\mathrm{IQR}, \qquad L_S = Q_3 + 1.5\,\mathrm{IQR}.$$

Luego, se marcó cada observación como candidata a atípica y se contó el total por va-

Tabla 3.1. Valores nulos detectados por clase

Variable	Positivos	Negativos
A1_Score	0	0
A2_Score	0	0
A3_Score	0	0
A4_Score	0	0
A5_Score	0	0
A6_Score	0	0
A7_Score	0	0
A8_Score	0	0
A9_Score	0	0
A10_Score	0	0
age	2	2
gender	0	0
jundice	0	0
austim	0	0
clase	0	0

riable:

$$o_i = \mathbf{1}\{\, x_i < L_I \vee \; x_i > L_S \,\}, \qquad N_{out} = \sum_{i=1}^{n} o_i.$$

donde:

- X : *Variable analizada.*
- x_i : *i-ésima observación.*
- n : *Tamaño de muestra.*
- Q_1, Q_3 : *Primer y tercer cuartil de X (equivalentes a los percentiles p_{25} y p_{75}).*
- IQR : *Rango intercuartílico $(Q_3 - Q_1)$.*
- L_I, L_S : *Límites inferior y superior para detectar atípicos.*
- $\mathbf{1}\{\cdot\}$: *Función indicadora (vale 1 si la condición es verdadera; 0 en otro caso).*
- $\vee$: *Disyunción lógica ("o").*
- N_{out} : *Número total de observaciones marcadas como atípicas.*

Al aplicar este algoritmo a todo el conjunto de datos, la variable `austim` *apareció con $N_{out} = 49$ datos atípicos.*

En otras palabras, el criterio IQR está pensado para distribuciones continuas; cuan-

do se aplica a una variable con dos niveles, el conteo de atípicos en realidad refleja el desbalance entre categorías (la clase minoritaria) y no la presencia de errores o valores extremos. Para verificarlo, se elaboró la dispersión de `austim` *(Figura 3.3) superponiendo los límites IQR. Todos los puntos se ubican exactamente en los dos niveles válidos de la variable, y ninguno traspasa los límites; por tanto, los "atípicos" reportados no son atípicos reales, sino un artefacto de aplicar IQR a una variable binaria.*

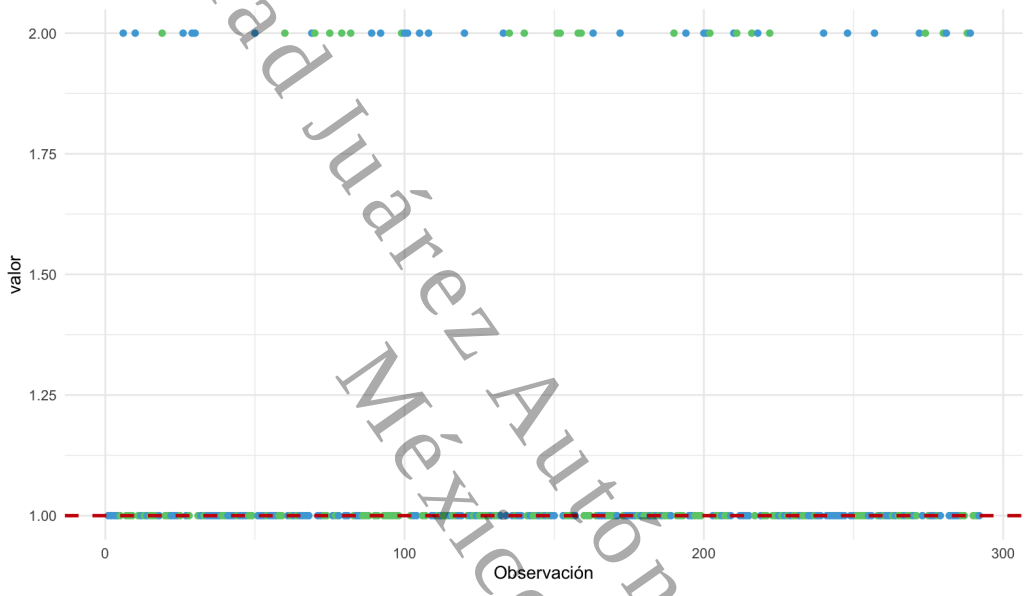

Figura 3.3. Gráfica de dispersión: puntos azules = clase 1, verdes = clase 2; la línea roja continua señala el umbral IQR, colapsado en 1 por la naturaleza binaria de la variable. Ningún punto se encuentra fuera del dominio $\{1, 2\}$; el criterio IQR no identifica atípicos reales.

Análisis estadístico de relaciones bivariadas

Se realizaron pruebas estadísticas para medir la relación entre cada variable predictora y la variable objetivo. Para ello se utilizaron dos métricas complementarias: la prueba Chi-cuadrada (X^2), que permite identificar asociaciones entre variables categóricas, y la Ganancia de Información, que mide cuánta información aporta cada atributo para reducir la incertidumbre en la predicción.

Como se aprecia en la Figura 3.4, los puntajes conductuales, en particular `A4_Score`*,* `A9_Score` *y* `A10_Score`*, mostraron asociaciones consistentes con la variable objetivo, lo que indica que tienen un papel relevante y podrían ser útiles en la construcción de mo-*

delos predictivos. En cambio, las variables de tipo demográfico, como `gender`, `jundice`, `austim` *y* `age`*, registraron valores muy bajos o cercanos a cero en la prueba Chi-cuadrada. Cabe señalar que únicamente* `age` *presentó una ligera señal de importancia bajo la métrica de Ganancia de Información.*

Además, se consideró la entropía como indicador de variabilidad y diversidad de los atributos. En este sentido, variables como `age` *presentaron mayor entropía, lo que refleja mayor heterogeneidad en sus valores y, por tanto, un mayor potencial de aportar información al modelo. Por el contrario, atributos como* `austim` *mostraron una entropía reducida, lo cual sugiere que sus valores son demasiado uniformes y que su aporte al modelo sería limitado.*

Estos resultados confirman que las variables de carácter conductual concentran la mayor capacidad explicativa frente a la variable objetivo, mientras que los factores demográficos apenas aportan información relevante. Esto sugiere que los modelos predictivos que se construyan a partir de estos datos estarán sustentados principalmente en patrones de comportamiento y no en características básicas como edad, género o antecedentes médicos, lo que refuerza la pertinencia de enfocar el análisis en los puntajes de evaluación psicológica.

Hallazgos en la exploración de datos

La exploración inicial permitió identificar patrones y depurar el conjunto de datos de manera más precisa. Se concluyó que algunas variables demográficas, como `jundice` *y* `austim`*, presentaban una relevancia estadística mínima y además contenían inconsistencias o distribuciones poco informativas, por lo que se consideró adecuado eliminarlas del análisis posterior.*

De igual forma, se detectaron valores faltantes en la variable `age`*. Dada su importancia relativa, se estableció la necesidad de imputarlos, proponiendo como estrategia más apropiada el uso de la mediana estratificada por clase, lo que permite conservar la coherencia de la distribución sin introducir sesgos significativos.*

Después de estas acciones de limpieza y depuración, el conjunto de datos quedó

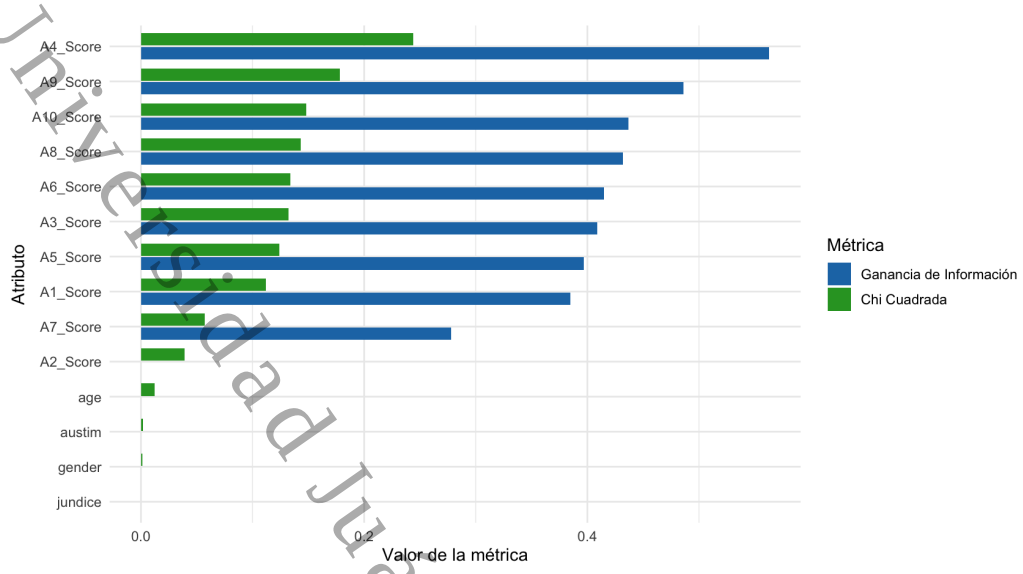

Figura 3.4. Comparación de importancia de atributos con dos métricas: barras azules = Ganancia de Información, barras verdes = Chi-cuadrada. Las barras se muestran lado a lado para cada variable (eje y), permitiendo ver que la variable *A4_Score* domina en ambas medidas, mientras que variables demográficas como *age*, *austim*, *gender* y *jundice* aportan prácticamente cero al modelo.

conformado por 292 observaciones y 13 variables.

3.2.2. Epileptic Seizure Recognition Data Set

El análisis exploratorio del conjunto de datos Epileptic Seizure Recognition Data Set fue conducido siguiendo el mismo proceso sistemático y ordenado que el dataset anterior: revisión de la estructura del dataset, detección de valores faltantes, identificación de posibles anomalías, análisis estadístico de relaciones bivariadas y, finalmente, una valoración preliminar del poder informativo de las variables. Esta secuencia permitió establecer un panorama claro sobre la calidad de los datos y su idoneidad para la fase posterior de modelado.

Inspección estructural

El conjunto de datos original estuvo conformado por un total de $11,500$ *observaciones y 180 variables. Estas se agruparon en tres categorías principales, como se muestra en la Tabla 3.2:*

- ***Grupo X**: incluye únicamente una variable de tipo numérica (`X`), que actúa como índice. No presentó valores faltantes.*
- ***Grupo X1 a X178**: conformado por 178 variables explicativas (`X1` a `X178`) de tipo entero, todas ellas completas, sin valores faltantes.*
- ***Clase**: corresponde a la variable objetivo (`clase`), definida como un factor binario tras la transformación aplicada (convulsión epiléptica = "Positivo", otros estados cerebrales = "Negativo"). Tampoco presentó valores faltantes.*

Cabe señalar que, originalmente, la variable `clase` estaba codificada en cinco categorías que representaban distintos estados cerebrales, siendo únicamente la categoría etiquetada como `1`, la asociada a eventos convulsivos. Para esta investigación, dicha clasificación fue simplificada a un esquema binario, lo que permitió enfocar el análisis en la distinción entre presencia y ausencia de convulsiones.

La ausencia total de valores faltantes en los tres grupos constituye una ventaja importante, ya que elimina la necesidad de imputación y facilita la preparación de los datos para etapas posteriores de depuración y modelado.

Tabla 3.2. Resumen estructural del conjunto de datos Epileptic Seizure Recognition

Categoría de variables	Cantidad	Tipo de dato	Descripción	Valores Faltantes
`X`	1	Numérico	Índice de registro	No
`X1` a `X178`	178	Entero	Variables explicativas (señales EEG)	No
`clase`	1	Factor (Binario)	Variable objetivo (Positivo/Negativo)	No

Detección de anomalías

Se aplicó el mismo criterio basado en el rango intercuartílico (IQR) que en el conjunto anterior, pero únicamente a las variables de señal `X1`–`X178` (numéricas y de alta

cardinalidad). Los umbrales se estimaron por variable usando todas las observaciones; la estratificación por clase se empleó después, solo para interpretar los resultados.

Se siguieron los pasos siguientes para cada variable X_j, con $j = 1, \ldots, 178$:

a. *Cálculo de cuartiles: Q_{1j} y Q_{3j}.*

b. *Rango intercuartílico: $IQR_j = Q_{3j} - Q_{1j}$.*

c. *Límites de Tukey: $L_{I,j} = Q_{1j} - 1.5\,IQR_j$ y $L_{S,j} = Q_{3j} + 1.5\,IQR_j$.*

d. *Etiquetado de atípicos: se marcó como atípica toda observación x_{ij} de X_j con $x_{ij} < L_{I,j}$ o $x_{ij} > L_{S,j}$ (equivalentemente, $x_{ij} \notin [\, L_{I,j},\, L_{S,j}\,]$).*

e. *Resumen de resultados:*

   a) *Por variable: número de registros de X_j fuera de $[\, L_{I,j},\, L_{S,j}\,]$.*

   b) *Por observación: para cada fila, número de variables X_j en las que el valor quedó fuera de los límites.*

Los resultados muestran una presencia extendida y casi uniforme de atípicos a lo largo de las señales: aparecen en prácticamente todas, con una media de atípicos por variable de 1254.23 y una desviación estándar de 136.96, lo que sugiere un comportamiento homogéneo entre X_1 y X_{178}.

La Figura 3.5 ilustra cuatro señales representativas (X_2, X_{50}, X_{100} y X_{150}); en ellas, los puntos correspondientes a las clases Negativo (azul) y Positivo (verde), junto con las líneas rojas discontinuas que marcan los umbrales basados en IQR, evidencian observaciones tanto por encima como por debajo de los límites, con un patrón que se repite entre variables, motivo por el cual no se graficaron las 178 señales. De forma complementaria, la Figura 3.6 sintetiza el conteo de atípicos por señal X_1–X_{178}, donde, pese a ciertas oscilaciones, la mayoría se concentra entre el intervalo de $1200 - 1350$, confirmando que ninguna señal está exenta de valores extremos.

Además del conteo por variable, se calculó para cada observación el número de señales que quedaron fuera de los límites (conteo por observación) y se comparó entre clases. En promedio, las observaciones de la clase Positivo (convulsión) presentaron

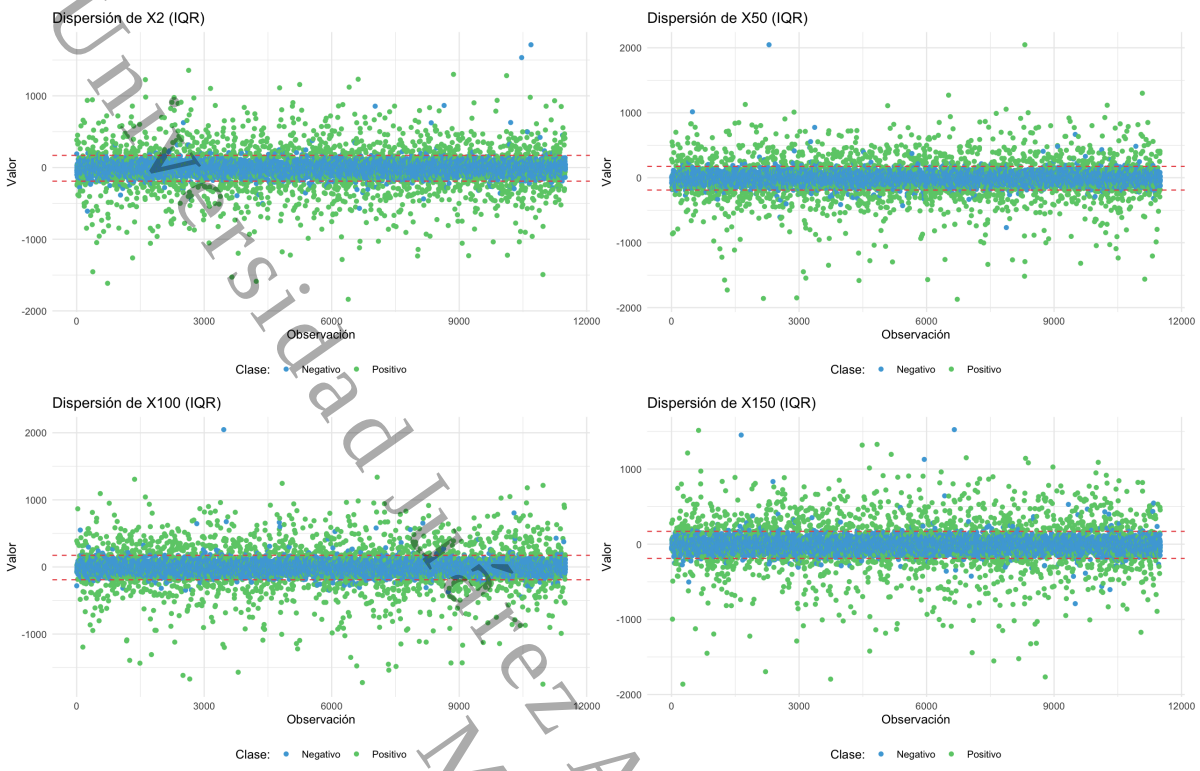

Figura 3.5. Dispersión de cuatro variables representativas (`X2`, `X50`, `X100`, `X150`) con límites IQR.

$\sim 7\times$ más atípicos que las de Negativo. Este contraste sugiere que los "extremos" detectados mediante IQR reflejan picos y transiciones propios del episodio convulsivo, más que errores de medición o ruido aleatorio.

Como resultado, no se eliminaron atípicos de forma automática: en señales fisiológicas los valores extremos pueden contener información clínica útil y borrarlos podría quitar potencial al modelo.

Cuando sea necesario mitigar su efecto en algoritmos sensibles a extremos, se recomienda un tratamiento robusto, por ejemplo, estandarización robusta (centrar en la mediana y escalar por el IQR, calculada solo con el conjunto de entrenamiento para evitar fuga de información. En modelos basados en árboles se pueden conservar los valores originales.

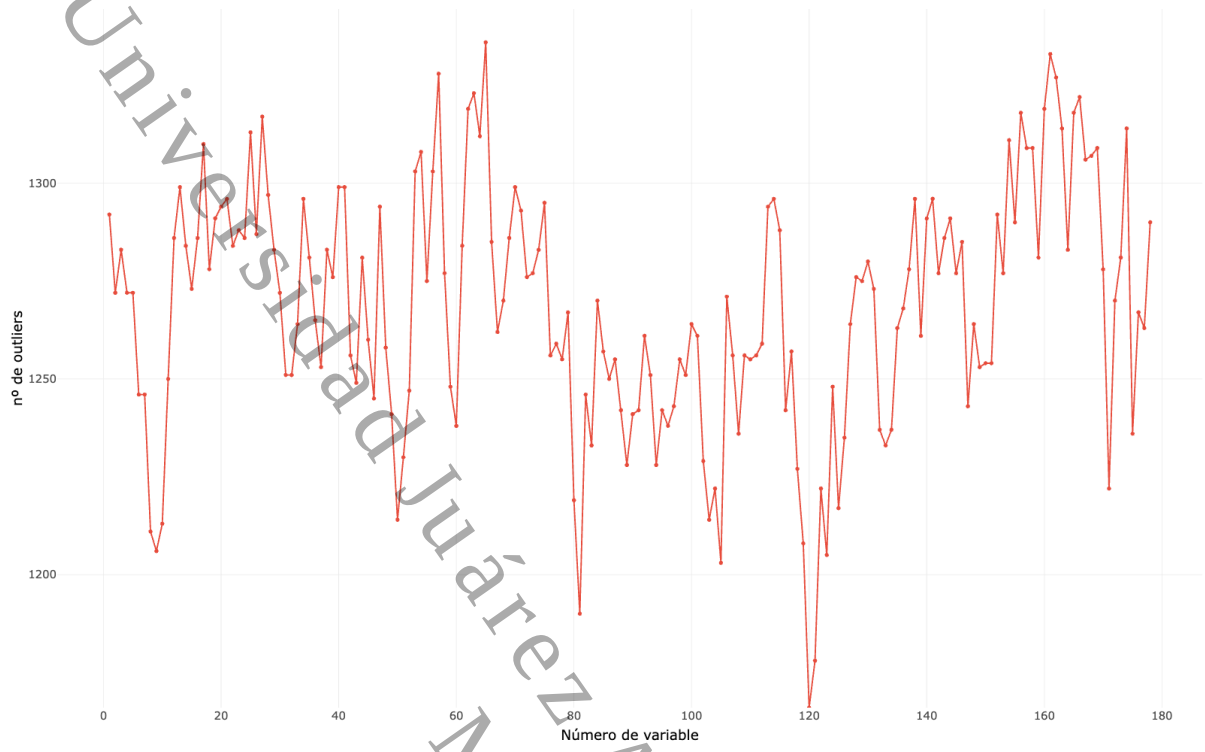

Figura 3.6. Número de valores atípicos detectados por variable de señal (`X1`–`X178`).

Análisis estadístico de relaciones bivariadas

Para cuantificar la asociación entre cada variable predictora y la variable objetivo se emplearon dos métricas complementarias: la prueba de independencia chi-cuadrada (tras discretizar las señales en cuantiles) y la ganancia de información. Los resultados, resumidos en la Tabla 3.3, muestran que la mayoría de las señales presentan un nivel de relevancia moderado y bastante uniforme (media de la puntuación reescalada de $\chi^2 = 0.655 \pm 0.070$*). Al comparar las mejores y peores puntuadas se observa que casi todas se agrupan alrededor de esa desviación estándar, sin diferencias marcadas; las únicas excepciones esperadas son la columna índice* `X` *y la propia* `clase`*, ambas con* $\chi^2 = 0$ *por no ser predictoras. En suma, salvo* `X` *y* `clase`*, las señales muestran relevancias muy similares y quedan elegibles para las etapas de modelado.*

Tabla 3.3. Resumen de los *top–bottom* cinco atributos según χ^2 y ganancia de información

Métrica	Mejores 5		Peores 5		Media ± DE	
	Variable	Valor	Variable	Valor	Media	Desv. Est.
χ^2	X38	0.681	X	0.000	0.655	0.070
	X128	0.681	clase	0.000		
	X39	0.680	X10	0.641		
	X127	0.679	X120	0.642		
	X111	0.677	X119	0.643		
Ganancia de información	X	0.722	X120	0.299	0.324	0.043
	clase	0.722	X119	0.301		
	X39	0.337	X10	0.302		
	X38	0.335	X137	0.304		
	X112	0.334	X148	0.305		
Entropía	X	13.489	clase	0.721	8.593	0.693
	X21	8.643	X30	8.578		
	X20	8.642.	X120	8.581		
	X127	8.637	X119	8.582		
	X128	8.637	X155	8.584		

Nota. Se muestran los cinco atributos con mayor y menor puntuación para cada métrica. Los valores se redondearon a tres decimales. Media y DE (desviación estándar) se calculan sobre la distribución completa de cada indicador.

Evaluación del potencial informativo mediante entropía

Se calculó la entropía univariada $H(X_j)$ de cada señal X_1–X_{178} a partir de su distribución empírica discretizada por cuantiles (agrupación en intervalos equiprobables). En concreto, cada variable se particionó en cinco intervalos de probabilidad similar (quintiles) y, como control de robustez, se verificó que el orden relativo de $H(X_j)$ se mantuviera al usar cuatro y seis intervalos (cuartiles y sextiles). Con las frecuencias de dichos intervalos se calculó $H(X_j)$ y se obtuvo el resumen de la Tabla 3.3. Los resultados muestran valores altos y bastante homogéneos (media 8.593 ± 0.693 en la escala utilizada), lo que sugiere que, en general, las señales presentan variabilidad interna suficiente y aportan información no redundante. Esta homogeneidad también se observa en los cinco mejores y peores casos de la tabla (por ejemplo, X_{21}, X_{20}, X_{127} y X_{128} entre las más altas; X_{30}, X_{120}, X_{119} y X_{155} entre las más bajas), cuyas diferencias respecto de la media son pequeñas. La columna índice `X` *exhibe la entropía más elevada por su naturaleza continua*

y rango amplio, mientras que `clase` *muestra un valor bajo por ser binaria; sin embargo, estos extremos no implican por sí solos capacidad predictiva. En este trabajo, la entropía se emplea como verificación de variabilidad y complementariedad (para descartar atributos degenerados o con información muy pobre); la priorización final se fundamenta en métricas bivariadas respecto de la etiqueta (prueba* χ^2 *y ganancia de información), no en* $H(\cdot)$ *de manera aislada.*

Hallazgos en la exploración de datos

Tras esta etapa de exploración, el experto determinó que ninguna variable presentaba datos faltantes ni inconsistencias estructurales significativas. La gran cantidad de anomalías encontradas en las observaciones positivas, representó un hallazgo importante que orientó la estrategia posterior de preprocesamiento y modelado, recomendando técnicas robustas como la winsorización o transformaciones específicas para preservar la integridad y el valor informativo del conjunto de datos.

El conjunto de datos final tras esta exploración mantuvo intactas las 11,500 observaciones, aunque se recomendó considerar la aplicación de técnicas robustas para el tratamiento de valores atípicos antes del modelado.

3.2.3. Diabetes Data set

El análisis exploratorio inicial del Diabetes Data Set, al igual que los conjuntos de datos anteriores, fue desarrollado bajo el mismo enfoque metodológico. Esta exploración, realizada por el experto humano, también abarcó desde la inspección estructural y la caracterización estadística, hasta la identificación de anomalías, el estudio de relaciones bivariadas y la evaluación de la variabilidad y relevancia predictiva de cada variable.

Inspección estructural

El punto de partida del análisis consistió en una inspección estructural minuciosa del conjunto de datos, conformado por 520 *observaciones y* 17 *variables clínicas, entre las que destaca la variable objetivo* `clase`*, encargada de distinguir entre sujetos con y sin*

diagnóstico de diabetes. Se verificó cuidadosamente la integridad del conjunto de datos: todos los atributos presentan información completa, sin valores faltantes en ninguna de las entradas. Esta ausencia de datos nulos no solo facilita la preparación y el procesamiento posteriores, sino que también aporta confianza en la calidad y fiabilidad de los registros analizados.

Asimismo, se revisaron y, en su caso, ajustaron los tipos de datos asignados a cada variable. De las 17 *variables,* 15 *eran de tipo binario, solo la variable* `Age` *era de tipo numérico y la variable* `clase` *era categórica. Se realizaron las transformaciones correspondientes para cada tipo de dato, asegurando la correcta interpretación por los algoritmos estadísticos y de aprendizaje automático empleados más adelante. Particularmente, la variable* `clase` *fue recodificada en un esquema binario claro y clínicamente relevante: 'Negativo' para sujetos sanos y 'Positivo' para quienes presentan un diagnóstico confirmado de diabetes. Este proceso ajustó las bases para una exploración y modelado consistentes, eliminando ambigüedades y asegurando evitar posibles fuentes de error desde la etapa inicial a las etapas posteriores.*

Respecto a la distribución de clases, se observó un moderado desequilibrio: el 61.5 % de las instancias corresponden a la clase 'Positivo' (diabetes) y el 38.5 % a la clase 'Negativo' (sin diabetes), como se muestra en la Figura 3.7. Esta proporción sugiere que los modelos de clasificación deben considerar el balance de clases para evitar sesgos hacia la categoría mayoritaria.

Estadística descriptiva

La fase de estadística descriptiva permitió obtener un panorama general de las características y la variabilidad presentes en el conjunto de datos. Inicialmente, se calculó la distribución de la variable objetivo (`clase`*), identificando un moderado desequilibrio: el 61.5 % de las instancias fueron clasificadas como positivas (diabetes), mientras que el 38.5 % correspondieron a la clase negativa (sin diabetes), como se muestra en la Figura 3.7.. Este desbalance es relevante, ya que puede influir en el desempeño de los modelos predictivos y, por ello, debe considerarse en las etapas posteriores del análisis.*

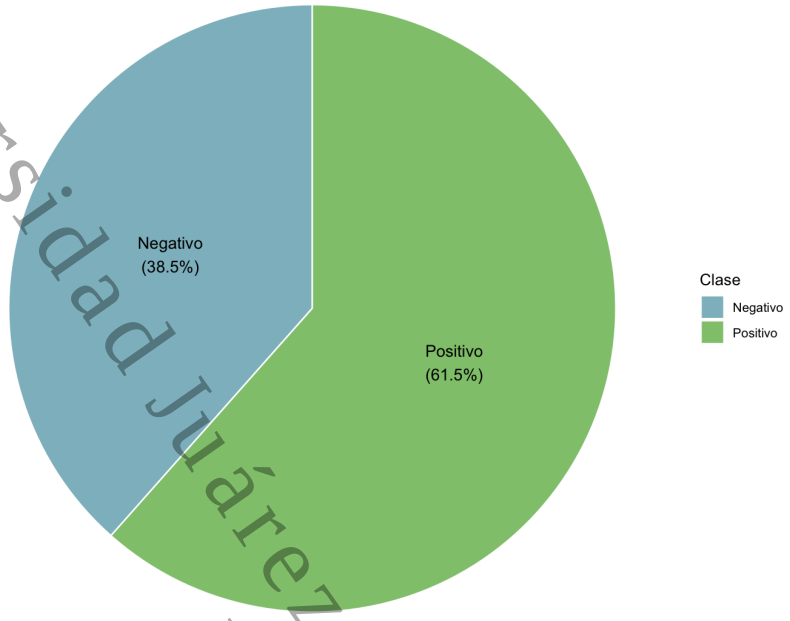

Figura 3.7. Gráfica de pastel de la variable `clase` de Diabetes Data Set.

Para cada una de las variables clínicas, se calcularon medidas de tendencia central (media, mediana) y dispersión (rango, cuartiles, desviación estándar), lo que permitió identificar valores típicos y posibles variaciones extremas en los datos. Por ejemplo, la variable `Age` *registró edades entre* 16 *y* 90 *años, con una mediana de edad de* 47.5 *años y cuartiles en* 39 *y* 57 *años, reflejando así una dispersión considerable en la población analizada.*

Las visualizaciones generadas, como la gráfica de pastel para la variable clase y los histogramas y boxplots para Age, facilitaron la interpretación gráfica de los datos y la identificación de patrones en la variable de interés. En particular, se observó que la distribución de edades es amplia y muestra cierta variabilidad, mientras que la variable clase presenta un desbalance que deberá considerarse en etapas posteriores del análisis.

Detección de anomalías

Para la identificación de observaciones atípicas en las variables numéricas, se utilizó el método del rango intercuartílico (IQR). Tal como se muestra en la Figura 3.8, la variable `Age` *identificó cuatro observaciones por encima del límite superior. No se encontraron valores atípicos en el extremo inferior.*

Sin embargo, considerando que la mediana de `Age` *es de* 47.5 *años y que la dispersión de edades se extiende hasta los* 90 *años, estos valores no deben interpretarse como errores, sino como casos menos frecuentes pero posibles dentro del contexto clínico de la diabetes.*

En las demás variables clínicas, la mayoría de tipo binario, no es apropiado aplicar la regla del rango intercuartílico para identificar valores atípicos, ya que sus valores sólo pueden ser 0 *o* 1*. Para variables como* `Genital.thrush`*,* `Irritability` *y* `Obesity`*, el conteo elevado de supuestos datos atípicos, responde a la frecuencia de la categoría positiva, no a la presencia de valores extremos en sentido estricto.*

Al analizar la distribución de registros con al menos un valor atípico por clase, se observó que la mayor proporción se encuentra en los sujetos con diagnóstico positivo de diabetes, lo que puede reflejar mayor variabilidad clínica en ese grupo.

Análisis de relaciones bivariadas

Para identificar cuáles atributos son más útiles para predecir la presencia de diabetes, se calcularon dos métricas: chi-cuadrada y ganancia de información. Ambas permiten medir la relación entre cada variable clínica y la variable objetivo tal como se muestra en la Figura 3.9.

En la prueba de chi-cuadrada, las variables `Polyuria` *y* `Polydipsia` *destacaron con los valores más altos (*0.66 *y* 0.65*, respectivamente), lo que indica una asociación fuerte entre estos síntomas y el diagnóstico de diabetes. Otras variables como* `Gender`, `sudden.weight.loss` *y* `partial.paresis` *también mostraron una asociación relevante, aunque en menor grado.*

En el análisis de ganancia de información, se observaron resultados consistentes:

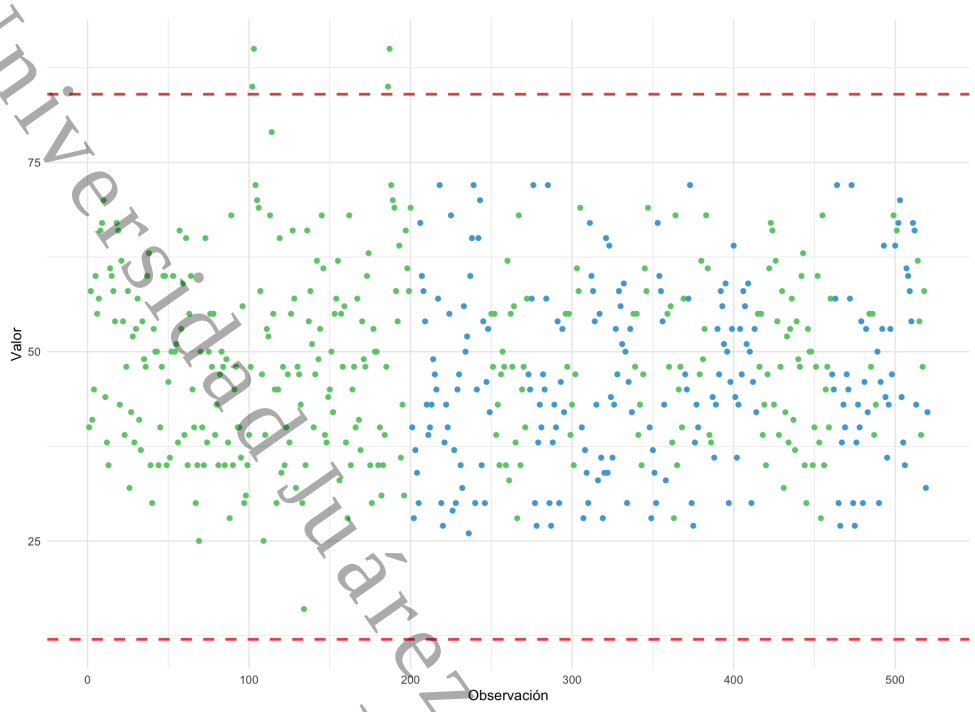

Figura 3.8. Dispersión de la variable `Age` con límites IQR

`Polyuria` (0.36) *y* `Polydipsia` (0.36) *nuevamente aparecieron como los atributos más informativos, seguidos por* `Age` (0.24) *y* `Gender` (0.16)*. Esto significa que estas variables aportan mayor cantidad de información útil para distinguir entre pacientes con y sin diabetes.*

Por otro lado, variables con valores bajos en ambas métricas, como `Itching`, `delayed.healing`, `Obesity` *y* `Genital.thrush`, *mostraron poca relevancia para la predicción. Es importante señalar que, aunque la obesidad es un factor de riesgo conocido para la diabetes, en este conjunto de datos no resulta un atributo discriminante debido a su baja variabilidad o distribución similar entre los grupos analizados. En resumen, ambas métricas permitieron priorizar los atributos más relevantes y, al mismo tiempo, identificar aquellos que aportan poca información al modelo.*

Evaluación de variabilidad con entropía

Se calculó la entropía para cada atributo con el fin de medir su diversidad. La variable `Age` *presentó el valor más alto de entropía* (5.26)*, lo que indica una mayor variabilidad*

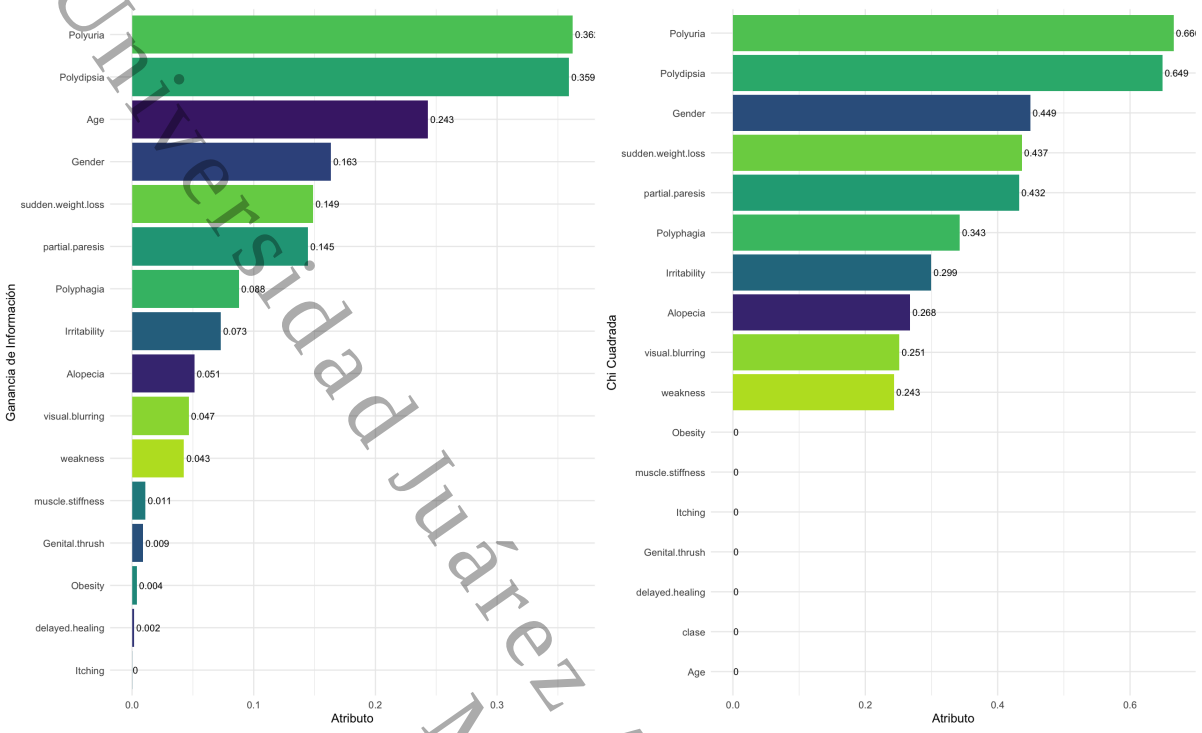

Figura 3.9. Gráficas de barras de Ganancia de Información (izquierda) y Chi-cuadrada (derecha)

en los datos de esta variable. En contraste, variables como `Obesity`, `Genital.thrush`, `Irritability` *y* `Alopecia` *mostraron menor variabilidad, lo que sugiere que contienen información redundante o limitada para la predicción del diagnóstico.*

Hallazgos en la exploración de datos

El análisis confirmó que el Diabetes Data Set no presenta valores nulos ni inconsistencias estructurales significativas. La mayoría de los valores atípicos se concentran en un pequeño grupo de variables y en la clase positiva. Las métricas de importancia y entropía permitieron identificar los atributos clave y fundamentar la posible eliminación o transformación de variables con baja utilidad predictiva para la siguiente fase de preprocesamiento.

3.3. Exploración de datos (IA Generativa)

3.3.1. Configuración del LLM y diseño del prompt para EDA reproducible

Con el objetivo de estandarizar y acelerar la Exploración de Datos (EDA), se diseñó un prompt específico para configurar un modelo de lenguaje grande (Large Language Model, LLM), en particular usando el modelo GPT-4o de OpenAI, a fin de que operara de manera interactiva y acotada al rol de "experto en ciencia de datos". La intención fue que el modelo funcionara como modo agente interactivo: primero solicitara información mínima indispensable al usuario y solo después generara un archivo R Markdown (RMD) completamente resuelto, sin placeholders, es decir sin marcador de posiciones. Todo esto para generar un archivo final listo para ser descargado y ejecutado en un entorno R con las librerías indicadas.

Este prompt fue diseñado para resolver un problema recurrente al utilizar LLMs en el análisis exploratorio de datos (EDA): los prompts genéricos suelen generar código basado en suposiciones, intentando adivinar rutas de archivos, tipos de datos o incluso la variable objetivo. Esto provoca lo que comúnmente se denomina "alucinaciones", un fenómeno en el que el modelo inventa información para completar vacíos cuando no cuenta con datos suficientes. Además, los modelos suelen intercalar placeholders que un usuario no técnico difícilmente puede reemplazar de manera adecuada. Para evitarlo, nuestro diseño exige que el modelo recopile cuatro datos esenciales del usuario (y opcionalmente un resumen del conjunto de datos) antes de generar la primera línea de código. Este enfoque reduce la ambigüedad, previene errores derivados de suposiciones y, en consecuencia, mejora significativamente la calidad del código final.

Rol y acotación contextual

Se instruyó explícitamente al LLM a "actuar como experto en ciencia de datos". Esta directiva de rol cumple dos funciones:

a. *Delimitar el alcance temático, restringiendo respuestas que se alejen de la tarea (salidas fuera de contexto).*

b. Fijar expectativas sobre el estilo de salida (claridad, orden y criterios propios de un EDA).

Encabezado del R Markdown y salidas

El prompt fija una cabecera YAML específica con título, autor, fecha dinámica y salida tipo `html_document` con tabla de contenido flotante, secciones numeradas y tema `readable`. Esta configuración garantiza un documento navegable y apto para revisión, con la capacidad de visualizar las gráficas interactivas, sin exigir edición posterior.

Contenido técnico solicitado al LLM (estructura del RMD)

El archivo RMD generado debía integrar las respuestas del usuario y producir un EDA ordenado en secciones, con código R limpio y comentado, y gráficos interactivos cuando corresponde.

El contenido solicitado fue:

1. Lectura de datos

- ***Carga de librerías:*** *Se importan las librerías necesarias para el análisis, como `dplyr`, `ggplot2`, `plotly`, `readr` y `tidyr`, entre otras.*
- ***Importación de datos:*** *Se lee el archivo usando la ruta y el tipo especificados, y el resultado se asigna a un DataFrame llamado* `datos`*.*
- ***Verificación de la carga:*** *Se muestra una vista preliminar de las primeras filas del DataFrame para confirmar que la importación fue exitosa.*

2. Análisis descriptivo general

- ***Dimensiones y estructura:*** *Se solicitó una inspección inicial del DataFrame para conocer sus dimensiones (número de filas y columnas) y la estructura de los datos, identificando el tipo de cada variable (numérica, categórica, entre otras).*

- ***Resumen de variables numéricas:*** *Se solicitó estadísticas descriptivas para las variables numéricas, incluyendo medidas de tendencia central (media, mediana), de posición (cuartiles) y de dispersión (desviación estándar).*
- ***Análisis de variables categóricas:*** *Para cada variable categórica, se solicitó la visualización de su distribución de frecuencias mediante un gráfico de barras interactivo, utilizando preferentemente la librería* `plotly` *o la combinación* `ggplot2` *+* `ggplotly()`*.*
- ***Organización del informe:*** *Con el fin de mejorar la legibilidad y el rendimiento del documento, se solicitó que cada una de las visualizaciones se encapsulara en un bloque de código independiente.*

3. Análisis de datos faltantes

- ***Cuantificación y localización:*** *Se solicitó un conteo de valores nulos (NA) por cada columna para determinar la magnitud de los datos faltantes.*
- ***Visualización de patrones:*** *Se solicitó que generara gráficos como un mapas de calor (heatmap) interactivo, utilizando preferentemente la librería* `plotly`*, para visualizar la distribución de los valores ausentes en el conjunto de datos. Este gráfico es clave para detectar si la ausencia de datos es aleatoria o si sigue un patrón específico.*

4. Detección de valores atípicos

- ***Metodología de Identificación:*** *Se especificó que se aplicaran métodos estadísticos para detectar valores atípicos en las variables numéricas. Principalmente, se utiliza el criterio del rango intercuartílico (IQR).*
- ***Visualización y reporte:*** *Adicionalmente se solicitó que generara un diagrama de caja y bigotes (boxplot) interactivo. Esta visualización permite una inspección gráfica inmediata de los outliers. Adicionalmente, junto a cada gráfico se reporta el conteo exacto de los valores atípicos identificados.*

5. Selección de características relevantes

- ***Análisis de correlación (variables numéricas):*** *Se solicitó que genera un mapa de calor (heatmap) interactivo para visualizar la matriz de correlaciones. El foco principal es identificar la fuerza y dirección de la relación lineal entre las variables predictoras y la variable objetivo, así como detectar posible multicolinealidad.*
- ***Ranking de características:*** *Como método alternativo o complementario, se construye una tabla que clasifica la importancia de cada característica según métricas estadísticas. La selección de la métrica depende del tipo de variable:*
  - *Para relaciones numéricas: Coeficientes de Pearson (lineal) o Spearman (monotónica).*
  - *Para relaciones con variables categóricas: Pruebas como Chi-cuadrado o métricas como la ganancia de información.*
- ***Presentación de resultados:*** *Se solicitó que todos los hallazgos se presenten en gráficos y tablas autoexplicativos e independientes, listos para una inspección clara que fundamente las decisiones del proceso de selección de características.*

6. Preparación para el modelado (recomendaciones)

- ***Transformación de variables:*** *Se solicitó que se especificaran las variables de tipo carácter (character) que deberán ser convertidas a un formato numérico en la etapa de preprocesamiento. Este paso es indispensable para que puedan ser utilizadas por los algoritmos de modelado.*
- ***Codificación de la variable objetivo:*** *Para los modelos de clasificación, fue fundamental asegurar que la variable objetivo sea declarada como tipo factor. Esta conversión garantizaría que los algoritmos la interpreten correctamente como una variable categórica con niveles definidos, evitando que sea tratada como una variable numérica continua.*

- ***Guía para el preprocesamiento:*** *Finalmente, se solicitó una muestra de nuevo de la estructura completa del* `Dataset`*. Esta vista consolidada serviría como una hoja de ruta (roadmap) para planificar y ejecutar de manera ordenada las transformaciones en la siguiente fase del proyecto.*

7. Consideraciones de implementación

- ***Calidad y claridad del código:*** *Se especificó que se debería escribir código limpio y legible, incluyendo comentarios en cada sección para explicar la lógica detrás de los procedimientos. Esto es fundamental para facilitar la colaboración y la mantenibilidad del proyecto.*
- ***Reproducibilidad garantizada:*** *Se especificó que el documento final debe ser autocontenido, lo que significa que debe poder ejecutarse de principio a fin en un entorno limpio sin generar errores. Este principio asegura que los resultados sean completamente reproducibles por terceros.*

Justificación técnica del enfoque

El diseño de esta herramienta se fundamenta en principios clave que garantizan su superioridad sobre enfoques genéricos. Se prioriza la reusabilidad al separar los parámetros del código, permitiendo su aplicación universal a distintos conjuntos de datos sin necesidad de reescritura. Para maximizar la robustez, el modelo solicita información crítica en lugar de adivinarla, previniendo así errores comunes.

El producto final es un informe RMD listo para ejecutar, eliminando la necesidad de edición manual por parte del usuario. Además, la interactividad de las visualizaciones con gráficas avanzadas construidas `plotly` *(o* `ggplotly()`*) aporta inspección dinámica en el navegador, útil para revisión guiada de categorías, faltantes y outliers. sumado a esto, la estructura modular del código a través de bloques y secciones separadas (chunks) asegura su legibilidad y facilita el mantenimiento. Finalmente, todo el análisis está alineado con el objetivo de clasificación, asegurando que cada paso del EDA aporte valor directo a la tarea de modelado.*

Figura 3.10. Flujo operativo para EDA asistida por LLM: (1) diseño del prompt; (2) espera activa hasta reunir ruta, variable objetivo, dimensiones y tipo de archivo; (3) generación de un R Markdown ejecutable para el EDA.

Este diseño establece un patrón reproducible para el Análisis Exploratorio de Datos (EDA). El proceso es interactivo en la fase de parametrización (véase la Figura 3.10), determinista en la organización del documento y consistente en los artefactos que genera (tablas, gráficos y diagnósticos). El prompt actúa como una plantilla inteligente que puede reutilizarse en diversos conjuntos de datos, eliminando la necesidad de reedición, preservando la consistencia metodológica y, finalmente, optimizando la eficiencia del flujo de trabajo.

3.3.2. Autistic Spectrum Disorder Screening Data for Children

Lectura de datos

El conjunto de datos, proveniente de un archivo de valores separados por comas (CSV), fue cargado en un DataFrame. Como verificación inmediata de que el archivo se leyó correctamente, se mostraron las primeras seis filas, que constituyen la cabecera del

conjunto de datos.

Resumen general

A continuación, el modelo generó estadísticas descriptivas para variables numéricas, revelando la distribución de puntajes del test AQ-10 infantil, edad y presencia de condiciones previas. Para las variables categóricas, como `gender`*,* `jundice`*,* `austim` *y la clase objetivo (*`clase`*), se produjeron gráficos de barras interactivos utilizando* `plotly`*, lo que permitió visualizar la frecuencia relativa de cada categoría con una experiencia de usuario enriquecida.*

Análisis de datos nulos

En cuanto a los valores faltantes, la IA identificó que la única variable con datos ausentes fue `age`*, con 4 valores nulos. Esta situación se representó mediante un mapa de calor interactivo (Figura 3.11) construido con* `ggplot2` *y habilitado para interacción con* `plotly`*, lo que facilitó la detección visual de las omisiones y su localización dentro del conjunto de datos*

Detección de anomalías

Para la detección de valores atípicos se construyeron diagramas de caja por variable numérica aplicando la regla del rango intercuartílico (IQR). La Figura 3.12 muestra el boxplot agrupado por variable, generado con `ggplot2` *y habilitado para interacción mediante* `ggplotly()`*. Se observan escalas heterogéneas: las puntuaciones* `A1_score`*–*`A10_score` *se concentran en un rango estrecho, mientras que* `age` *exhibe mayor dispersión.*

Identificando observaciones extremas en algunas de ellas. El modelo aplicó la metodología basada en el rango intercuartílico (IQR), generando visualizaciones agrupadas por variable mediante la función `ggplotly()`*.*

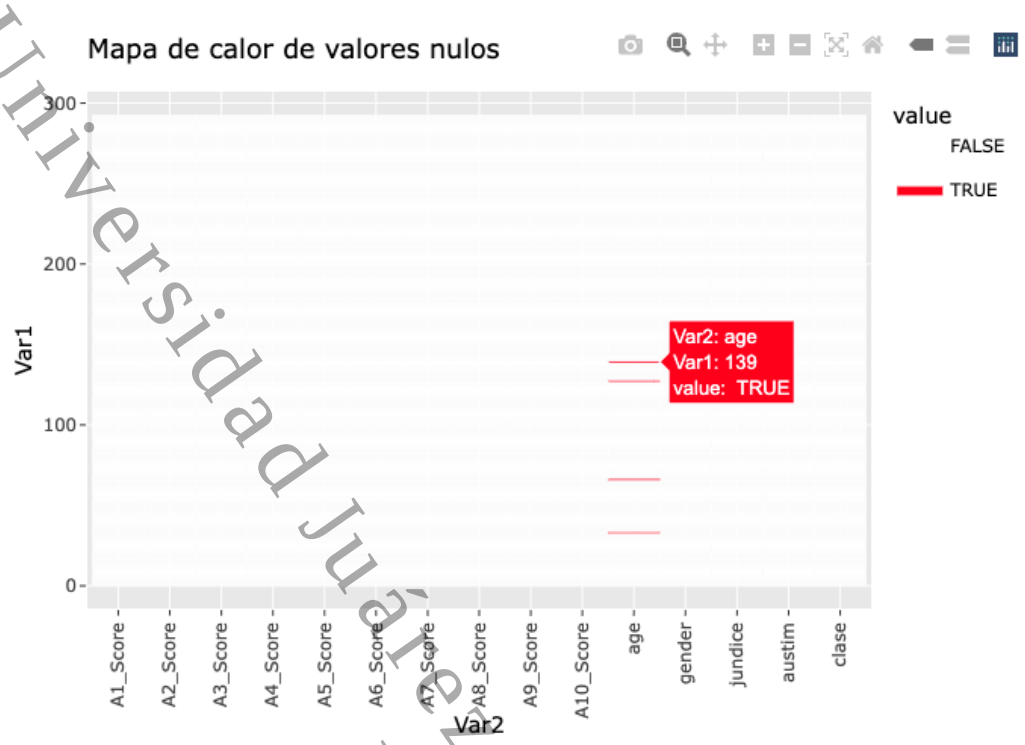

Figura 3.11. Mapa de calor de valores faltantes en *Autism-Child*. La única variable con omisiones es `age` (4 nulos).

Selección de características relevantes

La selección de características se abordó mediante el cálculo de la matriz de correlación de Pearson entre las variables numéricas, visualizada como mapa de calor (Figura 3.13) con `corrplot`. Esta inspección permite detectar redundancias (pares altamente correlacionados) y orientar la reducción de dimensionalidad. Aunque en esta etapa no se implementaron métricas como ganancia de información ni la prueba de chi-cuadrado, se propusieron como pasos siguientes para priorizar atributos en función de su relación con la variable objetivo

Preparación para el modelado

Finalmente, se sugirió la conversión de la variable objetivo `clase` a tipo `factor`, explicando su importancia para la etapa de modelado. Este tipo de recomendaciones demuestra la capacidad del modelo no solo para generar código, sino también para guiar al usuario en buenas prácticas de análisis de datos.

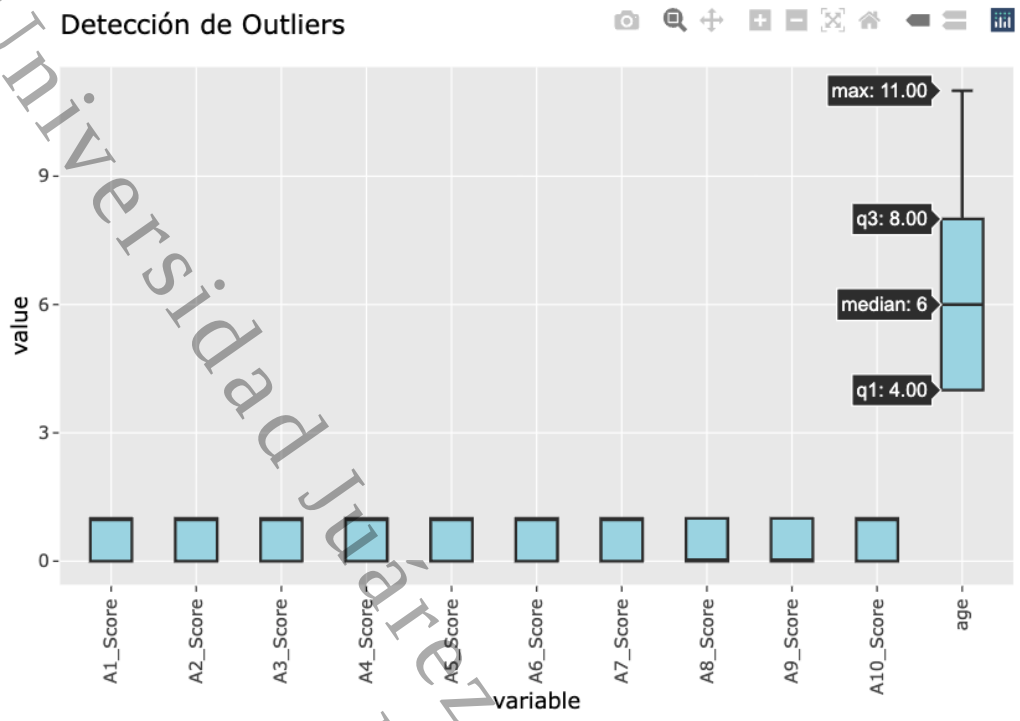

Figura 3.12. Diagrama de caja por variable numérica en *Autism-Child*. Las puntuaciones `A1_score`–`A10_score` se concentran en un rango acotado, mientras que `age` presenta mayor dispersión ($Q_1 = 4$, mediana$= 6$, $Q_3 = 8$, máximo$= 11$).

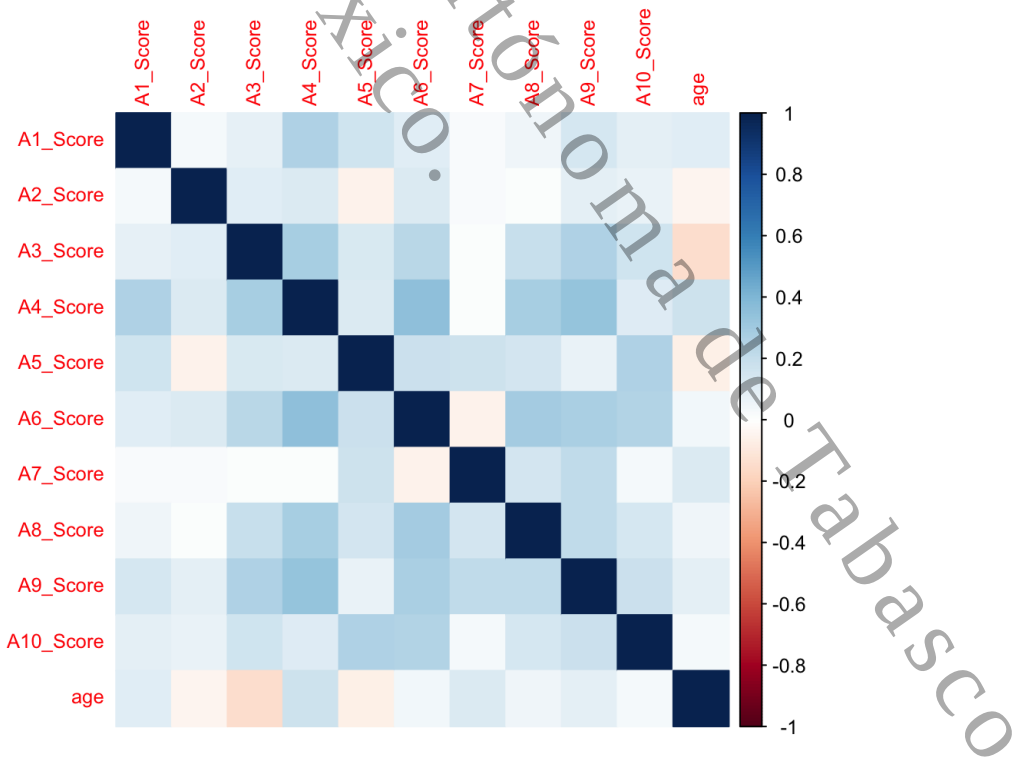

Figura 3.13. Matriz de correlación de Pearson entre variables numéricas en *Autism-Child*, representada como mapa de calor con `corrplot`. La escala va de -1 (correlación negativa) a 1 (correlación positiva), con diagonal unitaria.

En general, la IA generativa mostró un desempeño eficiente en la automatización del análisis exploratorio, generando código funcional, estructurado y con visualizaciones interactivas. No obstante, se evidenció que su efectividad depende de la claridad del prompt inicial y del acompañamiento humano para asegurar la validez de las decisiones analíticas propuestas.

3.3.3. Epileptic Seizure Recognition Data Set

Lectura de datos

El primer paso fue la carga y validación inicial de los datos. Se inspeccionó un extracto del conjunto, mostrándolo en un tibble (un formato de tabla moderno en R) de 6 filas y 180 columnas, lo que permitió verificar la correcta lectura del archivo. Esta validación de integridad confirmó que el conjunto de datos se cargó exitosamente en memoria y que su estructura era la adecuada para proceder con el análisis.

Resumen general

Como resumen general, se reportó que el conjunto de datos tenía una dimensión de 11,500 observaciones y 180 variables, con un resumen de su tipo de datos, donde la variable `X` *era de tipo carácter (identificador) y las variables restantes* `X1` *a* `X178` *eran de tipo numérico; asimismo, se listaron sumarios por variable (mínimos, cuartiles, medianas y máximos) que evidenciaron el rango amplio típico de señales EEG. Este resumen fue ejecutado correctamente y fue suficiente para entender la estructura de los datos antes del modelado.*

También, como se muestra en la Figura 3.14, la IA generó una gráfica de barras para la variable objetivo `clase`*. Este gráfico exhibió una distribución perfectamente uniforme: exactamente 2,300 observaciones en cada una de las cinco clases (1–5). Esto confirmó que, en el esquema original de cinco categorías, el conjunto estaba balanceado. Sin embargo, al reagrupar en el esquema binario utilizado en este estudio (clase 1 = convulsivo vs. clases 2–5 = no convulsivo), la misma distribución implicó un desbalance 1:4 (2,300 frente a 9,200 observaciones).*

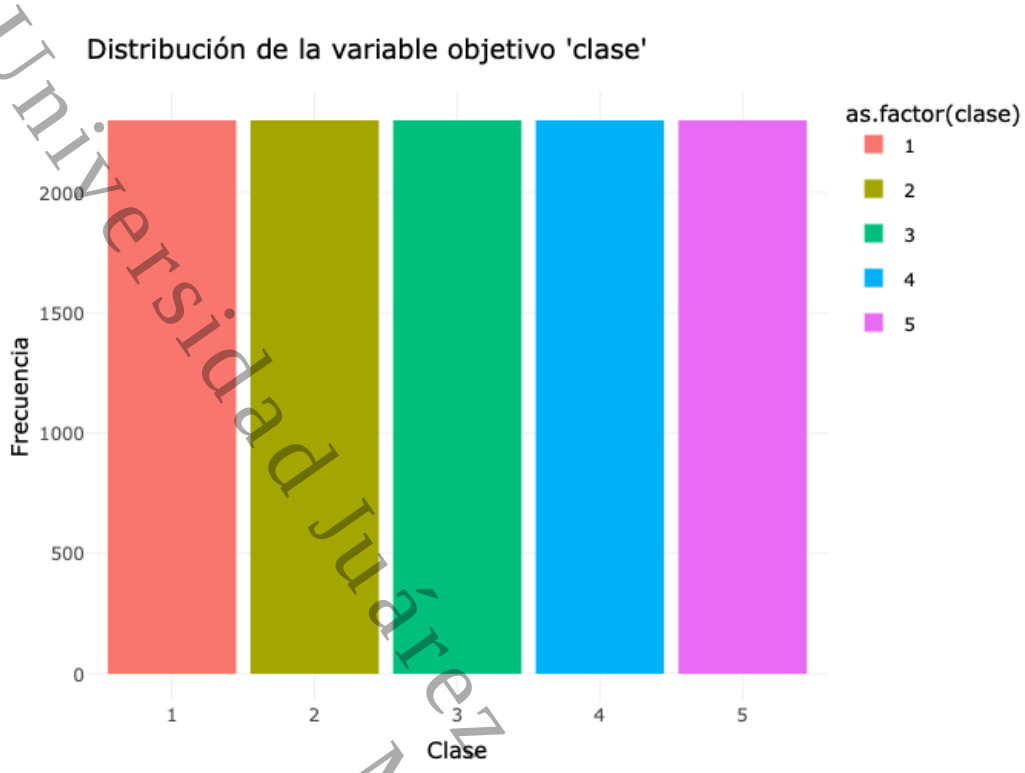

Figura 3.14. Distribución de la variable objetivo `clase` en el conjunto *Epileptic Seizure Recognition*: gráfico de barras con cinco categorías (1–5), cada una con exactamente 2,300 observaciones ($n = 11{,}500$). Imagen generada por la IA a partir de `as.factor(clase)`.

Análisis de datos nulos

Para el análisis y la detección de datos nulos, se cuantificaron explícitamente los datos faltantes y el resultado fue de 0 valores nulos en todo el conjunto de datos, lo cual se documentó de manera clara. El procedimiento fue correcto y suficiente.

Detección de anomalías

Para la detección de datos atípicos, el documento generado por la IA incluyó una sección de outliers con diagramas de caja para un subconjunto de variables representativas: `X`, `X1`, `X2`, `X3` *y* `X4` *(véase Fig. 3.15). Los diagramas de caja permitieron apreciar la mediana, la dispersión y posibles valores extremos en esas señales; no obstante, la selección abarcó solo 4 de las 178 variables numéricas del conjunto, por lo que la evidencia no fue exhaustiva a nivel global. Además,* `X` *correspondía a un identificador, por lo que su diagrama no resultó informativo para la detección de atípicos. Si bien la salida de la IA*

documentó indicios de valores extremos en las variables mostradas, no caracterizó su prevalencia ni su distribución en todo el conjunto.

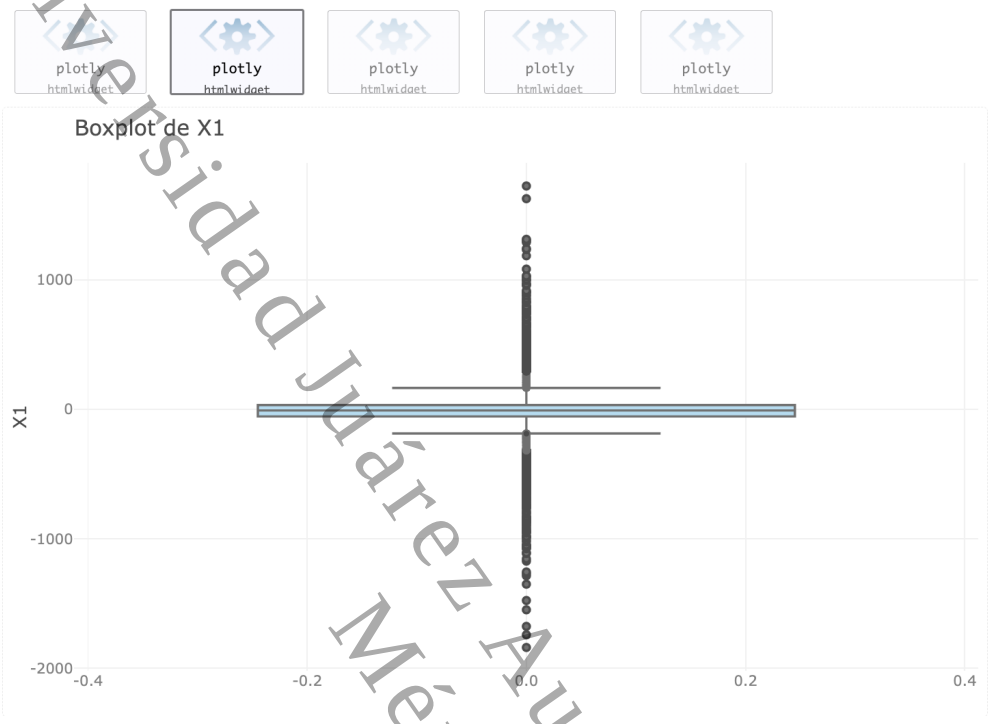

Figura 3.15. Diagramas de caja generados por la IA para variables seleccionadas (`X`, `X1`–`X4`) del conjunto *Epileptic Seizure Recognition*. Se visualizan mediana, rango intercuartílico y valores extremos por variable.

Selección de características relevantes

La IA presentó una matriz de correlación de Pearson como mapa de calor (como se muestra en la Figura 3.16), acotada a las primeras 30 variables numéricas (`X1`–`X30`). La variable `X` no apareció, previsiblemente por ser un identificador no numérico, lo cual fue correcto para este tipo de visualización. El uso de Pearson y de un mapa de calor simétrico con diagonal unitaria fue adecuado para mostrar, de forma compacta, la intensidad de las asociaciones lineales entre predictores. En la figura también se incluyó `clase` tras su codificación numérica; esto resultó válido como recurso de presentación homogénea, si bien su naturaleza categórica exigía cautela interpretativa. Dado que la imagen resumió solo un subconjunto del total de 178 señales, su alcance fue ilustrativo y no pretendía abarcar la totalidad del espacio de variables.

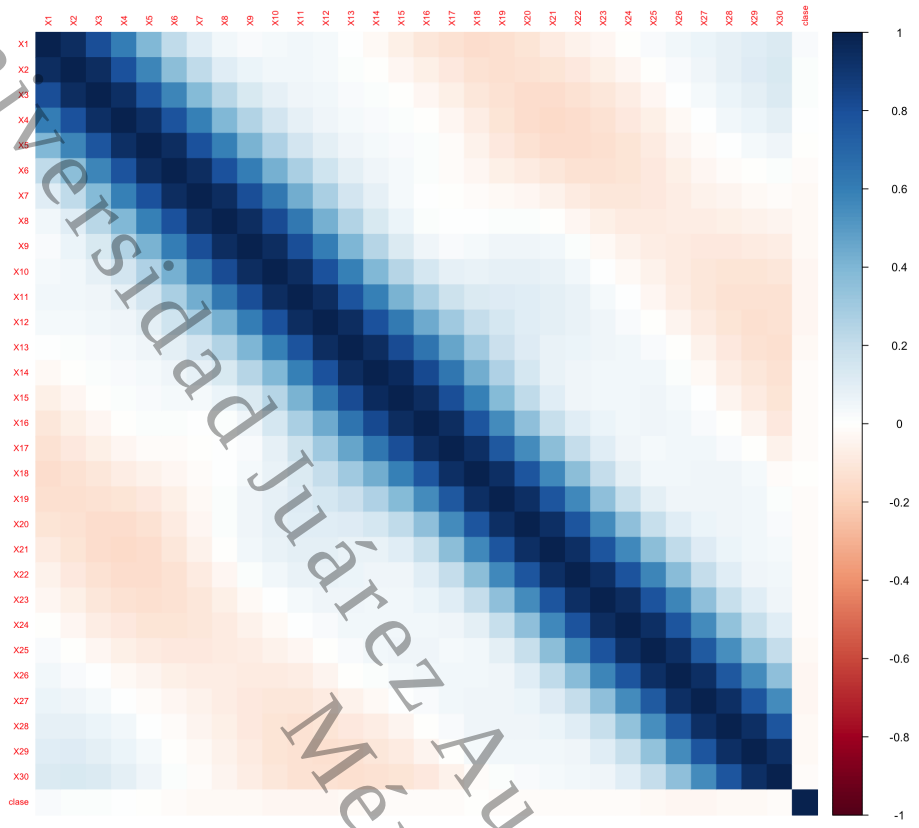

Figura 3.16. Mapa de calor de correlaciones de Pearson entre predictores numéricos (`X1`–`X30`) del conjunto *Epileptic Seizure Recognition*. Se omite `X` por no ser numérica. La escala va de -1 a 1 y la diagonal es unitaria.

Preparación para el modelado

En la etapa de sugerencias reportada por la IA, se efectuaron dos operaciones estrictamente descriptivas. Primero, se identificaron las columnas de tipo carácter: el resultado arrojó `TRUE` únicamente para la variable `X` *(identificador) y `FALSE` para* `X1`–`X178` *y* `clase`*, lo que confirmó que las señales EEG estaban almacenadas como variables numéricas. Segundo, para la variable objetivo se aplicó una conversión de tipo y se verificó que* `clase` *quedó representada como factor de cinco categorías (1–5) para fines de conteo y visualización. En ese marco, se sugirió omitir* `X` *de los resúmenes numéricos y tratar* `clase` *como un factor consistente.*

Finalmente, la IA registró la ausencia de valores faltantes en todas las variables. Para la variable objetivo, también mostró una distribución exactamente uniforme por clase (1–5): 2,300 observaciones por categoría sobre un total de $n = 11{,}500$. Al reagrupar en

el esquema binario empleado en este estudio (1 = convulsivo; 2–5 = no convulsivo), la misma partición se documentó como 1:4 (2,300 vs. 9,200). La evidencia se presentó mediante conteos y proporciones, tras convertir `clase` *a factor, y mediante un gráfico de barras; no se aplicaron ajustes ni técnicas de corrección, pues la sección se limitó a describir el estado de las clases y a dejar constancia para etapas posteriores.*

3.3.4. Diabetes Data Set

Lectura de datos

El tratamiento inicial de los datos comenzó con la importación del conjunto contenido en el archivo de datos. Esto se realizó de forma correcta, usando la función adecuada para tratar el tipo de archivo previamente especificado, en este caso de tipo csv. Como primer paso de validación, se efectuó una inspección dimensional que constató la presencia de 520 observaciones y 17 variables.

Posteriormente, la IA procedió a realizar la coerción y tipificación de datos para asegurar la consistencia requerida para el análisis subsecuente. En este proceso, la variable `Age` *se definió como de tipo numérico, mientras que las 16 variables restantes, incluida la variable objetivo* `class`*, se transformaron a un formato categórico (factor).*

Resumen general

Después, se ejecutó un resumen estructural del conjunto de datos resultante para confirmar que la lectura y la tipificación eran correctas. Este arranque comprobó que el archivo se leyó bien y que cada campo estaba en el tipo esperado para su análisis.

Una vez cargado correctamente el conjunto de datos, se procedió a realizar el Análisis Exploratorio de Datos. Se documentaron las dimensiones del conjunto, confirmando una dimensión estructural de 520×17*, y se produjo un resumen estadístico únicamente de la variable* `Age` *(ya que es la única variable de tipo numérica). Este resumen numérico arrojó medidas de tendencia central, como la media (*$\bar{x} = 48.03$*) y la mediana (*$Me = 47.5$*), así como medidas de posición y dispersión, incluido el primer cuartil (*$Q1 = 39$*), el tercer cuartil (*$Q3 = 57$*) y los valores mínimo (*16*) y máximo (*90*).*

De forma complementaria, para el análisis de las 16 variables de naturaleza categórica, se implementó un procedimiento automatizado para visualizar sus distribuciones de frecuencia. Mediante un bucle iterativo, se generó de manera sistemática un diagrama de barras para cada variable.

Dado que la inclusión de todas las gráficas resultantes excede los límites de este documento, se presenta la Figura 3.17 como un ejemplo representativo de este análisis. En ella se ilustra la distribución de la variable Obesity, mostrando la frecuencia absoluta de sus dos categorías ('Yes' y 'No'), lo que permite una evaluación inmediata de su desbalance.

```r
# Visualización de variables categóricas (excluyendo Age)
categoricas <- names(select(Dataset, -Age))

for (var in categoricas) {
  p <- ggplot(Dataset, aes(x = .data[[var]])) +
    geom_bar(fill = "steelblue") +
    theme_minimal() +
    labs(title = paste("Distribución de", var), x = var, y = "Frecuencia") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
  print(ggplotly(p))
}
```

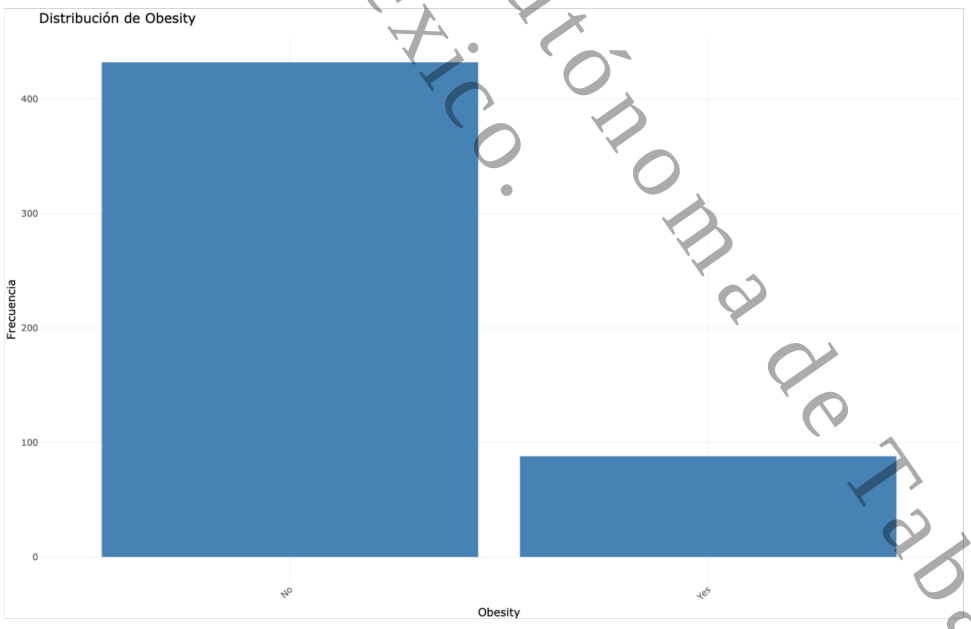

Figura 3.17. Ejemplo representativo de la exploración categórica en el *Diabetes Data Set*. La imagen corresponde a una de las 16 gráficas generadas automáticamente por la IA mediante un fragmento de código en bucle que recorre todas las variables categóricas excluyendo `Age`.

Como parte crucial de la validación de la calidad de los datos, se realizó una evalua-

ción de completitud en todo el conjunto. Se efectuó un conteo para verificar la existencia de valores nulos en cada variable, confirmando que ninguna de las 17 presentaba observaciones ausentes.

Análisis de datos nulos

Adicionalmente, se generó un mapa de calor de valores faltantes, que se mostró en la Figura 3.18, para proporcionar evidencia visual de que, efectivamente, no hubo omisiones. Este procedimiento fue correcto y dejó constancia clara de la ausencia de nulos en el conjunto de datos.

Figura 3.18. Mapa de valores faltantes del *Diabetes Data Set*. Se verificó la completitud de las 17 variables en 520 registros, confirmándose ausencia total de nulos (0 %). La visualización muestra que el 100 % de las celdas están observadas.

Detección de anomalías

Para el análisis de la detección de anomalías, dado que `Age` *era la única variable numérica, la IA construyó un diagrama de caja, una gráfica interactiva exclusivamente para esta variable. Esta decisión fue coherente con la estructura del conjunto (las demás variables eran categóricas) y permitió detectar visualmente valores atípicos en* `Age`*. En*

este caso, se mostraron 2 valores atípicos para las observaciones cuyos registros de edad fueron de 85 y 90 años.

Selección de características relevantes

Aunque en el prompt no se especificó que se utilizara Random Forest, la IA optó por implementarlo para producir un ranking de importancia de variables (véase Fig. 3.19). Esta elección resultó interesante, primero porque fue una decisión no especificada y, segundo, porque en este contexto la mayoría de las variables del conjunto eran categóricas y, por tanto, los enfoques basados en correlaciones lineales con la clase (como Pearson o Spearman) no eran adecuados para medir el aporte predictivo directo hacia una etiqueta binaria. Alternativas clásicas para variables categóricas (p. ej., chi-cuadrado o ganancia de información) habrían requerido supuestos y preparaciones adicionales; en cambio, el Random Forest permitió estimar la relevancia de cada predictor sin transformar el conjunto de datos en esta fase y sin alterar su contenido original. En términos de presentación, la salida cumplió el objetivo del EDA: ofrecer una vista preliminar, ordenada y comprensible, de qué variables parecían más influyentes, sin que esto implicara decisiones de modelado ni reducción de dimensionalidad en ese momento.

En la figura 3.19 se mostraron tanto el fragmento de código utilizado (entrenamiento del modelo y extracción de importancias) como el gráfico con las 15 variables mejor posicionadas. Se entendió como un recurso descriptivo para orientar la lectura y no como una validación definitiva del aporte de cada variable; su función aquí fue resolver el requerimiento de "identificar variables relevantes" de manera compatible con la naturaleza de los datos y dejar evidencia reproducible de cómo se obtuvo dicho ranking.

3.3.4.1. Preparación para el modelado

Se imprimió nuevamente la estructura del conjunto de datos para constatar los tipos: `Age` *como numérica; el resto, como factores; y* `class` *como factor de dos niveles ("Negative", "Positive"). La IA, además, dejó constancia de que la variable objetivo se manejó como factor para clasificación. Estas acciones fueron descriptivas y propias de la etapa*

```r
# Evaluación de importancia con Random Forest (variable objetivo: class)
modelo <- train(class ~ ., data = Dataset, method = "rf", importance = TRUE)
var_imp <- varImp(modelo)
plot(var_imp, top = 15)
```

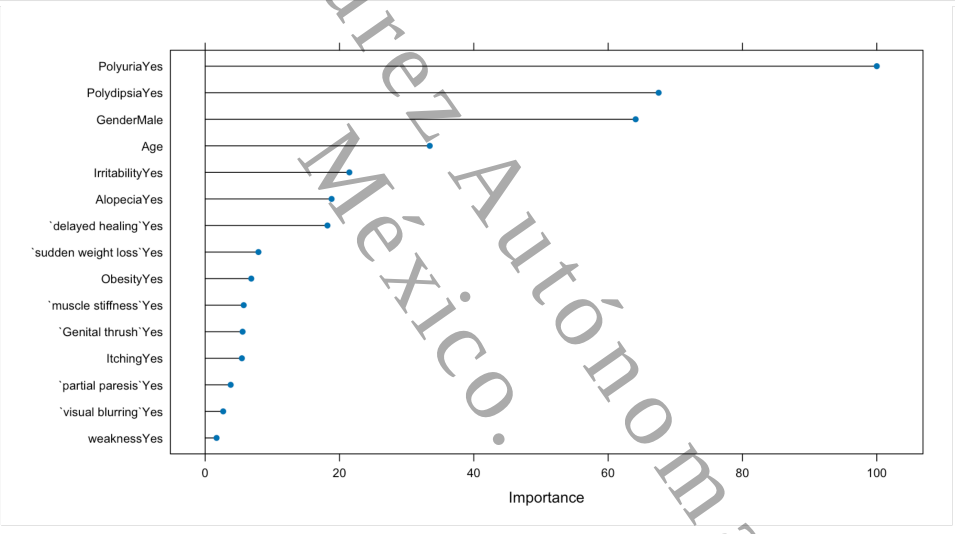

Figura 3.19. Ranking de importancia de variables en el *Diabetes Data Set* generado por la IA con un modelo de Random Forest. La figura incluye el fragmento de código empleado y el gráfico con las 15 variables mejor posicionadas. Se usa como recurso descriptivo en EDA para presentar una vista preliminar de relevancia sin modificar el dataset.

exploratoria

El reporte generado por la IA incluyó una sección de cierre ("Consideraciones Adicionales") en la que se indicó que el EDA proporcionó la base para la siguiente etapa de preparación y modelado; no se aplicaron modificaciones sobre el conjunto de datos en esta fase.

3.4. Preparación de los datos (experto Humano)

La preparación de datos, desde la perspectiva del experto humano, tuvo como propósito transformar el conjunto original en un insumo coherente, consistente y listo para modelarse, sin perder trazabilidad respecto de lo observado en la exploración. Todas las decisiones que siguieron se fundamentaron en la evidencia de la etapa previa de exploración de datos (tipos de variables, ausencia/presencia de nulos, posibles atípicos, balance de la clase, entre otros) y se ejecutaron con tres principios: integridad (no introducir sesgos ni eliminar información útil), reproducibilidad(procedimientos documentados y con control de aleatoriedad definiendo una semilla aleatoria) y sin fuga de información (ninguna transformación aprendió del conjunto de prueba).

El flujo de trabajo se organizó de la siguiente manera:

a. ***Definición de roles de las variables.*** *Separar identificadores y metadatos de los predictores; validar el tipo de cada campo (numérico o categórico) y la codificación de la variable objetivo*

b. ***Calidad y consistencia.*** *Verificar por última vez datos faltantes, rangos plausibles (valores que tiene sentido en la realidad) y valores fuera de dominio; cuando aplique.*

c. ***Escala y atípicos.*** *Estandarizar o normalizar únicamente las variables numéricas que lo requirieron y definir el manejo de atípicos (por ejemplo, escalamiento robusto o winsorización.*

d. ***Codificación de categóricas.*** *Elegir un esquema claro y estable (binarización/indicadores, codificación ordinal cuando correspondiera), garantizando consistencia entre entrenamiento y prueba*

e. ***Balance de la clase.*** *Registrar el reparto de la variable objetivo, sin modificar el conjunto original de prueba.*

f. ***Selección/derivación de características.*** *Elaborar, cuando fuera pertinente, variables derivadas o reducir dimensionalidad, siempre como parte de un pasos reproducibles ajustados en entrenamiento y validado por validación cruzada.*

g. ***Partición y control de aleatoriedad.*** *Dividir el conjunto en entrenamiento y prueba con estratificación por la clase y semilla fija; a partir de este punto, el conjunto de prueba quedó bloqueado para cualquier decisión. Toda calibración (escalas, codificadores, seleccionadores) se ajustó exclusivamente con datos de entrenamiento.*

El resultado de esta etapa son dos objetos de trabajo: un conjunto de datos de entrenamiento con un documento (`rmarkdown`*) que contiene las transformaciones documentadas y reproducibles y un conjunto de datos para prueba reservado para la evaluación final.*

3.4.1. Autistic Spectrum Disorder Screening Data for Children

Esta sección se documentó, de forma verificable, las transformaciones aplicadas al conjunto Autistic Spectrum Disorder Screening Data for Children hasta la partición en conjuntos de entrenamiento y prueba. El objetivo fue dejar los datos en un estado consistente para el modelado, registrando cada decisión y su efecto sobre la estructura del conjunto de datos.

Establecimiento de tipos y recodificación de la clase *Se homogeneizaron los tipos de datos y se recodificó la variable objetivo:* `clase` *se renombró a dos niveles Negativo/Positivo, y se declaró como factor, tras un paso de tipificación general del conjunto.*

Eliminación de variables *Se eliminaron las columnas* `jundice` *y* `austim`*. Tras esta operación, el objeto resultante (*`datafinal`*) quedó con 292 observaciones y 13 variables.*

Imputación de valores faltantes *Se constató la existencia de cuatro valores nulos en la variable* `age`*. Para orientar la imputación, se calcularon medias, medianas y modas por clase; las medianas de* `age` *fue el valor de 6 años tanto en la clase Positiva como en la Negativa. Con base en ello, se imputó* `age` *con el valor de 6 años, con lo cual el conjunto de datos quedó sin nulos.*

Estado posterior a las transformaciones *Se verificó nuevamente la dimensión y se registró el conteo de variables/observaciones antes y después del preprocesamiento: Original tenía una dimensión de 21 variables por 292 observaciones, y al final, la dimensión del Preprocesado tuvo como resultado 13 variables y 292 observaciones.*

Partición en entrenamiento y prueba *Con semilla fija (123), se realizó una división de proporción 2/3 estratificada por la clase mediante la función* `createDataPartition`*, produciendo dos nuevos subconjuntos de datos nombrados como* `trainingdata` *y* `testingdata`*. A partir de este punto, el conjunto de prueba quedó reservado para la evaluación final.*

3.4.2. Epileptic Seizure Recognition Data Set

Esta sección describe, paso a paso, las modificaciones efectivamente aplicadas al conjunto de datos y los artefactos generados para dejarlo listo para su uso posterior.

Tipos de datos y recodificación de la variable objetivo *Se homogenizaron los tipos de dato con una transformación general que convirtió caracteres a factores y, a continuación, factores a numéricos en las variables predictoras. Después,* `clase` *se recodificó a un esquema binario con etiquetas* `"Positivo"` *y* `"Negativo"`*, y quedó declarada como factor. Estas operaciones se registraron de forma explícita en el documento.*

Integridad y completitud *Se verificó la ausencia de valores faltantes a nivel global en el conjunto; no se detectaron nulos. Esta constatación se documentó dentro del flujo de conocimiento de datos.*

Eliminación del identificador *Se eliminó la columna* `X` *por tratarse de un identificador que no aporta información predictiva directa. Tras esta depuración, el conjunto de datos se redujo a 11,500 filas y 179 variables.*

Escalado y centrado de predictores *Con el fin de estandarizar magnitudes entre señales, se aplicó centrado y escalado (*`scale()`*) a todas las variables numéricas, manteniendo* `clase` *sin transformar y reinsertándola al final. El resultado se consolidó en* `datafinal\_scaled`*.*

Revisión de valores atípicos sobre datos escalados *Se repitió la detección de atípicos (regla IQR) sobre la parte numérica ya escalada, y se reportaron conteos por variable mediante tablas de apoyo. Esta revisión dejó trazabilidad de los umbrales inferiores/superiores y del número de incidencias por señal.*

Selección de atributos relevantes *Segun el calculo de rankings por entropía, chi-cuadrado y ganancia de información. Entre los resultados ilustrativos donde se listaron los mejores/peores valores por métrica, se decidió preservar a la mayoria de columnas debido a que los valores resultantes, no implicaban diferencias significativas.*

Estado final *El flujo deja un conjunto depurado sin* `X`*, con* `clase` *en formato binario tipo factor, y con los predictores numéricos centrados y escalados en* `datafinal\_scaled`*.*

3.4.3. Diabetes Data Set

La fase de preparación del conjunto Diabetes Data Set fue relativamente sencilla en comparación con otros datasets analizados, debido a la limpieza inicial de los datos y la estructura bien definida del archivo. Este conjunto está compuesto por 520 instancias y 17 atributos clínicos, entre los cuales se encuentran variables continuas, categóricas y la variable objetivo relacionada con el diagnóstico de diabetes.

Una primera inspección reveló que el dataset no contenía valores nulos, lo cual fue confirmado mediante funciones de conteo por columna. Este hallazgo permitió omitir la

etapa de imputación, facilitando una transición directa hacia el análisis exploratorio y el modelado.

No obstante, se detectaron valores extremos (outliers) en algunas variables numéricas, particularmente en `insulin`*,* `glucose` *y* `BMI`*. Aunque estos valores no fueron eliminados de forma inmediata, fueron documentados y marcados como posibles candidatos para transformación en etapas posteriores, en función del comportamiento de los modelos de clasificación.*

Para garantizar la consistencia de los tipos de datos, se transformaron las variables categóricas en factores. En particular, la variable de salida fue convertida a un factor binario con los niveles `Positive` *y* `Negative`*, lo cual es un requerimiento técnico para la mayoría de los algoritmos supervisados en R.*

Por último, se aplicó una normalización mediante estandarización z-score a todas las variables numéricas, con el fin de uniformar las escalas y prevenir sesgos en modelos sensibles a las magnitudes, como redes neuronales, máquinas de vectores de soporte y KNN. Esta estandarización permitió mejorar la estabilidad numérica de los algoritmos y facilitó la interpretación de resultados en etapas posteriores del proceso.

La preparación realizada por el experto humano permitió asegurar que el conjunto estuviera en condiciones óptimas para su uso en modelos predictivos, maximizando la calidad de los datos sin alterar su estructura esencial.

3.5. Preparación de los datos IA Generativa (IA Generativa)

El objetivo de esta sección es explicar con precisión cómo se obtuvo, con apoyo de una IA generativa, un archivo R Markdown que automatiza la Preparación de los Datos. El proceso inicia con el reporte en PDF de la Etapa 1 (Exploración de Datos) y con un prompt diseñado para delimitar el contexto, solicitar información crítica al usuario y, únicamente después de recibirla, generar el código reproducible. Cada componente del prompt cumple una función concreta y se incluye para evitar supuestos, mantener trazabilidad y alinear el preprocesamiento con los hallazgos exploratorios.

Propósito del prompt *El prompt establece el alcance y la secuencia de trabajo. Se utiliza la instrucción "Actúa como experto en ciencia de datos" para delimitar el dominio de la IA (con ello se restringe la generación al contexto de preparación de datos, evitando respuestas genéricas o fuera de tema). Además, el prompt exige una interacción previa para recolectar parámetros clave antes de producir el R Markdown (esto previene colocar marcadores vacíos y reduce ambigüedades).*

Elementos que la IA recibe y produce

a. *Insumo principal: el PDF de la Etapa 1 (Exploración de Datos) con hallazgos sobre valores faltantes, valores atípicos, variables constantes y otros incidentes. Este documento guía las acciones de limpieza y transformación (así las decisiones se basan en evidencia previa).*

b. *Insumo operativo: la ruta o nombre del archivo del conjunto de datos. Se solicita explícitamente para garantizar reproducibilidad y evitar confusiones con versiones distintas.*

c. *Preferencias del usuario: criterios para tratamiento de nulos, manejo de valores atípicos, transformaciones y proporciones de partición. Se capturan de forma explícita (esto alinea el preprocesamiento con el objetivo analítico y con restricciones del dominio).*

d. *Salida: un archivo R Markdown con código en R, comentarios y visualizaciones, además de los archivos* `dataset_train.csv` *y* `dataset_test.csv` *(esto deja un registro ejecutable y exporta subconjuntos listos para modelado).*

Interacción previa obligatoria (y por qué) *Antes de generar código, la IA hace seis preguntas. Cada una existe para cerrar un riesgo metodológico concreto.*

a. *Confirmar la ubicación o el nombre del conjunto de datos (evita cargar una fuente incorrecta y preserva trazabilidad).*

b. *Solicitar un resumen de hallazgos del PDF de la Etapa 1 (asegura que el preprocesamiento responda a evidencia real como porcentajes de nulos, presencia de valores atípicos y variables con varianza casi nula).*

c. *Definir la estrategia para valores nulos (imputación o eliminación por filas o columnas; esta decisión balancea pérdida de información y posible sesgo por imputación).*

d. *Definir la estrategia para valores atípicos (detección por IQR o z scores y tratamiento mediante eliminación, winsorización o transformación; esto limita la influencia de observaciones extremas en algoritmos sensibles a la escala).*

e. *Indicar transformaciones adicionales (codificación de variables categóricas mediante one hot encoding o conversión a factor, escalado o normalización, corrección de tipos, eliminación de variables irrelevantes; con ello se mejora compatibilidad algorítmica y estabilidad numérica).*

f. *Establecer la proporción de partición entrenamiento y prueba (por ejemplo, 70/30 u 80/20, junto con una semilla; esto garantiza evaluación honesta y reproducibilidad).*

Solo después de tener estas respuestas, la IA genera el R Markdown. Esta condición existe para que el documento final no contenga campos vacíos ni supuestos no declarados.

Estructura del *R Markdown* que se genera *El archivo inicia con un encabezado YAML que define título, autoría, fecha y salida* `html_document` *con tabla de contenido y secciones numeradas (esto facilita navegación y auditoría). Luego, el contenido se organiza en apartados con propósito claro:*

- *Lectura del conjunto de datos (carga de librerías como* `dplyr`, `tidyr`, `readr`, `ggplot2`, `plotly` *y, cuando aplica,* `caret`, `rsample` *o* `caTools`*; lectura del archivo a un objeto* `Dataset`*; vista de primeras filas y estructura para confirmar tipos y dimensiones).*
- *Tratamiento de valores nulos (imputación numérica mediante media o mediana, imputación categórica mediante moda o una categoría "Desconocido", o eliminación*

por umbrales; se reportan conteos y porcentajes antes y después para cuantificar impacto).

- *Manejo de valores atípicos (detección con IQR y z scores; aplicación de la estrategia elegida; visualización con diagramas de caja y distribuciones para evidenciar cambios en dispersión y sesgo).*
- *Transformaciones y limpieza adicional (codificación de variables categóricas con one hot encoding o factores, escalado o normalización de continuas, corrección de tipos, fusión de niveles raros, eliminación de variables irrelevantes o con varianza casi nula; se muestra la nueva estructura del conjunto de datos).*
- *División en subconjuntos de entrenamiento y de prueba (uso de* `caret`, `rsample` *o* `caTools`*; fijación de semilla; en clasificación, se propone estratificación para conservar la distribución de la variable objetivo).*
- *Exportación y resumen final (guardado de* `dataset_train.csv` *y* `dataset_test.csv`*; impresión de dimensiones y conteo de variables por tipo; confirmación de rutas de salida).*

Criterios de decisión y justificación *Las decisiones se justifican con base en objetivos de calidad de datos y compatibilidad con modelado.*

- *Imputación frente a eliminación (la imputación conserva tamaño muestral a cambio de introducir supuestos; la eliminación evita supuestos pero puede reducir potencia estadística; por ello se reporta el impacto de cada opción).*
- *Detección y tratamiento de valores atípicos (IQR y z scores son criterios estándar para identificar extremos; la winsorización, la transformación y la eliminación controlan su influencia y se elige según el objetivo de interpretabilidad y la sensibilidad del algoritmo).*
- *Codificación, escalado y normalización (la codificación con one hot encoding evita ordinalidad ficticia; el escalado y la normalización reducen dominancia por magnitud y mejoran el desempeño de métodos que dependen de distancias o gradientes).*

- *Partición y semilla (la proporción entrenamiento y prueba define un compromiso entre ajuste y evaluación; la semilla vuelve el procedimiento replicable; la estratificación en clasificación preserva la distribución de la variable objetivo).*

Resultado y transición *El resultado es un R Markdown parametrizado que documenta y ejecuta la Preparación de los Datos de forma reproducible (incluye visualizaciones y reportes antes y después de cada paso). Con esta base, las subsecciones 3.5.1, 3.5.2 y 3.5.3 describen en detalle lo obtenido para tres conjuntos de datos específicos: Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set (en cada caso se reportan decisiones, transformaciones efectivas, evidencia visual del antes y después y la confirmación de los archivos de entrenamiento y prueba).*

3.5.1. Autistic Spectrum Disorder Screening Data for Children

La preparación automatizada del conjunto de datos Autism-Child se realizó utilizando el modelo GPT-4o, a través de un prompt estructurado que solicitó previamente información clave al usuario antes de generar el código. Esta interacción permitió adaptar el archivo R Markdown a las características específicas del dataset y a las decisiones analíticas tomadas en la etapa exploratoria.

Una vez proporcionada la ubicación del archivo `Autism-Child-Data.csv`, se confirmaron los principales hallazgos: presencia de valores faltantes en la variable `age`, ausencia de nulos en otras columnas, y distribución binaria en variables categóricas como `gender`, `jundice`, `austim` y `clase`. El usuario especificó que los valores nulos debían ser imputados por la mediana y que los valores atípicos serían conservados, documentados pero no eliminados. También se acordó convertir variables categóricas a factores y aplicar escalado a las variables numéricas, dividiendo posteriormente el conjunto en 70 % entrenamiento y 30 % prueba.

El modelo generó automáticamente un documento RMD con las siguientes secciones:

- ***Lectura del dataset**: Se cargaron las bibliotecas necesarias (`dplyr`, `readr`, `ggplot2`,*

`plotly`, `tidyr`) y se verificó visualmente la estructura de los datos mediante la función `head()`.

- ***Tratamiento de valores nulos**: Se aplicó la imputación por mediana a la variable `age`, empleando código generado dinámicamente que recalculó las estadísticas solo para esta columna. Se mostró un resumen actualizado para verificar que los valores nulos fueron eliminados correctamente.*
- ***Transformación de variables**: Las variables categóricas fueron convertidas a factores, en particular la variable objetivo `clase`. Se realizó un escalado de las variables numéricas mediante estandarización z-score. Esta transformación se documentó con un resumen estructural del nuevo dataset.*
- ***División del conjunto de datos**: El modelo incluyó código para particionar los datos en subconjuntos de entrenamiento y prueba utilizando la función `createDataPartition()` del paquete `caret`. Los subconjuntos fueron exportados como `dataset_train.csv` y `dataset_test.csv`, garantizando así su disponibilidad para la etapa de modelado.*

El proceso automatizado permitió obtener un script funcional, comentado y reproducible que prepara de manera adecuada el conjunto Autism-Child para tareas de clasificación. Si bien el modelo generó código robusto, su efectividad dependió directamente de la claridad de las decisiones proporcionadas por el usuario. Esto subraya la importancia de una colaboración asistida entre IA y humano durante el proceso de preparación de datos.

3.5.2. Epileptic Seizure Recognition Data Set

La preparación automatizada del conjunto de datos Epileptic Seizure Recognition se realizó con el modelo GPT-4o mediante un prompt interactivo que primero solicitó información esencial y, con base en ella, generó un archivo R Markdown reproducible. La interacción previa se centró en confirmar la ubicación del archivo `data.csv`, registrar los hallazgos clave de la etapa exploratoria y acordar criterios de preprocesamiento acordes

con dichos hallazgos. Este orden evita supuestos, mantiene trazabilidad y alinea cada transformación con la evidencia inicial.

A partir del resumen provisto, se estableció lo siguiente: el conjunto de datos contiene 11,500 observaciones y 180 variables, no presenta valores faltantes, concentra valores extremos en múltiples variables numéricas con rangos entre aproximadamente `-1800` y `+2000`, todas las variables son numéricas salvo `clase` que funge como objetivo, existe desbalance de clases cercano a 4 a 1 y no hay variables constantes. Con estos elementos, se definieron acciones que privilegian mantener la información disponible y documentar los efectos de cada paso, a fin de facilitar decisiones posteriores en modelado.

El documento R Markdown generado incluyó las siguientes secciones:

- ***Lectura y verificación inicial**: Carga de bibliotecas (`dplyr`, `tidyr`, `readr`, `ggplot2`, `plotly`, y `caret` para la partición) e importación de `data.csv`. Se validaron dimensiones y tipos mediante `head()` y `str()` para confirmar 11,500 filas, 180 variables y la naturaleza categórica esperada de `clase` tras su conversión.*

- ***Valores nulos**: Al no detectarse valores faltantes se omitieron rutinas de imputación. Se conservó un bloque de verificación que imprime conteos de nulos por variable como control de calidad en futuras ejecuciones.*

- ***Manejo de valores atípicos**: Dada la presencia de valores extremos en numerosas variables numéricas y con el fin de no descartar posibles patrones relacionados con eventos convulsivos, se optó por conservarlos en esta fase, documentando su distribución con criterios IQR y z scores. Se generaron diagramas de caja y distribuciones antes y después del diagnóstico para evidenciar su impacto en la dispersión. Esta decisión aplaza cualquier intervención más agresiva para la etapa de modelado, donde el algoritmo seleccionado y su sensibilidad a la escala orientarán si conviene winsorizar, transformar o excluir.*

- ***Transformaciones y estructura**: La variable objetivo `clase` fue convertida a `factor` y se aplicó estandarización tipo z score a las variables numéricas. Esta transformación se justifica por la amplitud de rangos observada y por la conveniencia de*

estabilizar magnitudes para métodos sensibles a escala y distancia. Se imprimió la estructura actualizada del conjunto de datos para verificar tipos y número de variables tras el escalado.

- ***Desbalance de clases**: Se reportó la distribución de `clase` y se incluyó una nota técnica que recomienda abordar el desbalance en el modelado mediante estrategias como ponderación de clases, muestreo o umbrales de decisión, según el algoritmo empleado. En la preparación se garantizó que las particiones respetaran la proporción observada.*
- ***División del conjunto de datos**: Se realizó una partición estratificada en 70 % entrenamiento y 30 % prueba con `createDataPartition()` del paquete `caret` (se fijó semilla para reproducibilidad). Esta elección asegura evaluación honesta del desempeño y preserva la proporción 4 a 1 de la variable objetivo en ambos subconjuntos. Los resultados se exportaron como `seizure_train.csv` y `seizure_test.csv`.*

El proceso asistido por IA produjo un script claro, comentado y orientado a la evidencia, que deja listo el conjunto Epileptic Seizure Recognition para su uso en modelado. La ausencia de valores nulos simplificó la limpieza, mientras que la presencia de valores extremos y el desbalance de clases se trataron mediante diagnóstico, estandarización y partición estratificada. La efectividad del preprocesamiento se sustenta en la correspondencia directa entre hallazgos exploratorios y acciones implementadas, y en que cada decisión quedó documentada para su revisión y ajuste en etapas posteriores.

Diabetes Data Set

La preparación del conjunto Diabetes Data Set mediante inteligencia artificial generativa se realizó utilizando el modelo GPT-40, a través de un prompt que permitía generar dinámicamente un archivo R Markdown (RMD) personalizado. Antes de producir el código, la IA interactuó con el usuario para recopilar información clave: nombre del archivo, hallazgos previos (por ejemplo, ausencia de valores nulos, presencia de outliers), y las estrategias preferidas para el preprocesamiento.

Dado que este conjunto de datos no contenía valores faltantes, la IA omitió cualquier procedimiento de imputación, lo cual fue verificado mediante código de inspección previa generado automáticamente. En cuanto a los valores atípicos, identificados en variables como `insulin`, `glucose` y `BMI`, se optó por conservarlos sin alteración, priorizando mantener la información original para evaluación posterior en los modelos.

El código generado por el modelo GPT-4o incluyó las siguientes etapas:

- ***Lectura de datos**: Importación del archivo correspondiente con `readr` y visualización de la estructura mediante `head()` y `str()`. Se confirmaron las dimensiones del dataset y el tipo de cada variable.*
- ***Tratamiento de outliers**: Si bien no se eliminaron, la IA generó visualizaciones (boxplots interactivos con `ggplotly()`) para permitir su monitoreo. Se destacó que la decisión de intervenir o no sobre ellos se podría posponer para fases posteriores del modelado.*
- ***Transformaciones adicionales**: Se convirtió la variable objetivo a tipo `factor` y se aplicó estandarización z-score a todas las variables numéricas. Estas transformaciones se realizaron con funciones bien comentadas, facilitando la reproducibilidad y la revisión manual del código por parte del usuario.*
- ***División de datos**: El conjunto preprocesado fue particionado en 70 % entrenamiento y 30 % prueba, utilizando la función `createDataPartition()` del paquete `caret`. Los subconjuntos fueron exportados en formato `.csv` como `diabetes_train.csv` y `diabetes_test.csv`.*

El resultado fue un archivo completamente funcional, estructurado y modular, que automatiza la preparación del conjunto Diabetes Data Set de forma eficiente. Aunque el conjunto presentaba menos complejidades técnicas que otros datasets, el uso de IA permitió generar un pipeline de preprocesamiento robusto, adaptable y directamente utilizable en la etapa de modelado. Este caso reafirma el potencial de la inteligencia artificial generativa como herramienta complementaria en flujos de trabajo analíticos estructurados.

3.6. Modelado (Experto Humano)

Para cada uno de los tres conjuntos de datos, el analista humano llevó a cabo la fase de modelado utilizando R. Se implementó un proceso sistemático que incluyó validación cruzada estratificada y la comparación de múltiples algoritmos de clasificación, con el objetivo de identificar el modelo de mejor desempeño para cada problema. En todos los casos, se consideraron las mismas ocho técnicas de modelado: árbol de decisión, regresión logística, k-vecinos más cercanos (KNN), máquina de vectores de soporte (SVM), red neuronal, bosque aleatorio, árbol potenciado (boosted tree) y un modelo de ensamble heterogéneo. Para garantizar la consistencia, estos modelos se configuraron con hiperparámetros fijos o previamente optimizados de forma similar en todos los conjuntos: por ejemplo, el árbol de decisión se entrenó con `rpart` *usando un parámetro de complejidad* `cp = 0.25`*, el modelo KNN usó* `k = 16` *vecinos, el SVM empleó un núcleo radial con costo* `C = 5`*, la red neuronal se definió con* `size = 1` *(una neurona oculta) y* `decay = 0.1`*, el bosque aleatorio utilizó* `ntree = 500` *árboles y* `mtry = 2` *variables por división, y el árbol potenciado se entrenó con un grado de interacción fijo (por ejemplo, profundidad 1) y una tasa de aprendizaje moderada (control* `c = 0.5` *en* `caret`*).*

Finalmente, se construyó un modelo de ensamble combinando las predicciones de varios de los clasificadores anteriores mediante voto mayoritario. En todos los casos, la variable respuesta es binaria, por lo que para los modelos que emiten probabilidades (Árbol de Decisiones, KNN, SVM, Red Neuronal y Regresión Logistica), se utilizó un umbral de 0.5 para definir la clase positiva. Durante el entrenamiento, se aplicó validación cruzada (por ejemplo, de 5 pliegues) para obtener estimaciones robustas del desempeño de cada modelo y guiar la selección de hiperparámetros en caso necesario. Las métricas principales recopiladas fueron la precisión (en este contexto, equivalente al valor predictivo positivo de la clase positiva), la exactitud balanceada (balanced accuracy), la sensibilidad (recall o tasa de verdaderos positivos) y la especificidad (tasa de verdaderos negativos). Estas métricas permiten evaluar no solo la proporción global de aciertos, sino también el comportamiento del modelo en cada clase, algo crucial sobre todo cuando las clases están desbalanceadas. Al término del modelado, cada modelo entrenado se

guardó en un archivo `.rds` (por ejemplo, `decision\_tree_model.rds`, `svm\_model.rds`, *etc.) para su posterior uso en la etapa de validación con el conjunto de prueba reservado. A continuación, se describen los detalles particulares y consideraciones tomadas para cada conjunto de datos durante el modelado por el experto humano.*

3.6.1. Autistic Spectrum Disorder Screening Data for Children

En el caso del conjunto Autism-Child, el analista humano exploró múltiples algoritmos de clasificación con el fin de detectar de manera precisa el Trastorno del Espectro Autista (ASD) en niños. Dado que el dataset es relativamente pequeño y no presentó valores perdidos tras la fase de preprocesamiento, se procedió directamente al entrenamiento de los modelos descritos anteriormente utilizando la muestra de entrenamiento. Se empleó validación cruzada estratificada para estimar el desempeño de cada modelo de forma consistente, asegurando que la proporción de niños diagnosticados con ASD se mantuviera en cada pliegue. Cada modelo se ajustó con los hiperparámetros previamente mencionados (por ejemplo, el árbol de decisión con `cp = 0.25`*, el KNN con* `k = 16`*, etc.), y se compararon las métricas de desempeño promedio. Gracias al tamaño manejable del conjunto de datos y a un cuidadoso preprocesamiento, el experto anticipó que algunos algoritmos podrían lograr una clasificación casi perfecta. Tras el entrenamiento, los modelos resultantes fueron almacenados en discos (`.rds`) para su validación posterior con datos nuevos.*

3.6.2. Epileptic Seizure Recognition Data Set

Para el conjunto de datos Epileptic Seizure Recognition, que contiene 11 500 instancias con 180 variables numéricas cada una, el modelado requirió consideraciones adicionales debido a la alta dimensionalidad y al desbalance de clases (aproximadamente una proporción 4:1 entre la clase mayoritaria y la minoritaria). El analista humano siguió un protocolo riguroso para entrenar los modelos, manteniendo la coherencia con los experimentos de los demás conjuntos pero introduciendo técnicas para mitigar el impacto del desbalance. Los mismos ocho algoritmos de clasificación mencionados previamen-

te fueron evaluados en este dataset, asegurando la reproducibilidad en la comparación. A continuación, se detallan los aspectos clave del protocolo seguido y las motivaciones detrás de él:

Protocolo de evaluación y justificación. *Se utilizó validación cruzada estratificada (con k pliegues) para entrenar y validar los modelos, de modo que cada partición preservara aproximadamente la proporción original de la variable objetivo. Esto fue fundamental para obtener estimaciones de desempeño más estables y evitar sesgos debidos al desbalance. Dado el rango muy diferente de escalas entre las 180 variables numéricas, se aplicó una estandarización (normalización Z) previa del conjunto de entrenamiento, con el fin de homogeneizar las magnitudes y prevenir que los predictores de mayor varianza dominaran indebidamente ciertos modelos (especialmente relevante para KNN, SVM y Redes Neuronales).*

En cuanto al desbalance de clases, se implementaron dos estrategias complementarias durante el modelado:

a. *Ponderación de clases en algoritmos que lo permiten (por ejemplo, asignando mayor peso a la clase minoritaria en la función de costo de SVM o en el cálculo de la pérdida de la regresión logística), y*

b. *Técnicas de muestreo en el conjunto de entrenamiento (sobre-muestreo de la clase minoritaria o sub-muestreo de la mayoritaria) dentro de cada pliegue de validación, cuando fue pertinente.*

Estas medidas buscaban mejorar la detección de la clase minoritaria (ataques epilépticos) sin alterar la evaluación externa, ya que la validación cruzada permaneció estrictamente estratificada y las métricas empleadas (especialmente la exactitud balanceada) reflejan explícitamente el desempeño en ambas clases.

Cada modelo se entrenó con los mismos hiperparámetros base definidos (por ejemplo, $cp = 0.25$ para el árbol, $C = 5$ para SVM). Las métricas de desempeño se calcularon a partir de las predicciones en los pliegues de validación, prestando especial atención a la exactitud balanceada como indicador principal, dado que combina la sensibilidad y la

especificidad y, por lo tanto, es menos influenciada por el predominio de una clase. Al finalizar el entrenamiento, los modelos ajustados se guardaron en archivos `.rds` *para su posterior evaluación con el conjunto de prueba independiente.*

Cabe mencionar que, en un conjunto de datos con estas características, era de esperar que los modelos que incorporan mecanismos de regularización o métodos de conjunto pudieran sobresalir (por ejemplo, SVM con kernel no lineal, bosques aleatorios o el ensamble heterogéneo), ya que la complejidad del espacio de atributos y la presencia de valores atípicos suelen penalizar a modelos más sencillos. En cualquier caso, la selección final del modelo para implementación se basaría en las métricas cuantitativas obtenidas bajo el mismo esquema de validación, asegurando una comparación justa entre enfoques.

3.6.3. Diabetes Data Set

El conjunto de datos Diabetes Data Set, compuesto por 520 instancias y 17 variables clínicas por paciente, se utilizó para desarrollar modelos de clasificación binaria orientados a predecir el diagnóstico de diabetes (positivo o negativo). Este dataset, obtenido tras la etapa de preprocesamiento, no contenía valores faltantes y presentaba una estructura bien definida de variables (edad, índices de salud, resultados de pruebas de laboratorio, etc.), por lo que el analista aplicó un pipeline de modelado estándar muy similar al de los demás casos. Un pipeline se entiende como una secuencia ordenada y automatizada de pasos (selección de características, partición de datos, escalamiento, entrenamiento y evaluación) que garantiza coherencia metodológica, reduce errores manuales y facilita la comparación entre modelos.

Tras estandarizar las variables numéricas y confirmar que la distribución de la variable objetivo no estuviera altamente desbalanceada (en este caso, la proporción de pacientes diabéticos contra los no diabéticos era manejable), se procedió a entrenar los ocho modelos establecidos. Nuevamente, se empleó una validación cruzada estratificada para entrenar y comparar los algoritmos, obteniendo métricas de precisión (valor predictivo positivo), exactitud balanceada, sensibilidad y especificidad en cada caso. Todos los

modelos usaron la misma configuración de hiperparámetros previamente definida (sin requerir ajustes especiales para este conjunto).

En general, durante la etapa de modelado, el desempeño preliminar de los modelos en la validación interna fue significativamente superior al azar, indicando que existían patrones útiles en las variables clínicas para predecir la diabetes. Una vez completado el entrenamiento, se exportaron los modelos entrenados en archivos `.rds`*, quedando listos para la etapa de evaluación externa sobre el conjunto de prueba reservado.*

3.7. Modelado (IA Generativa)

Para contrastar con el enfoque manual, también se empleó un modelo de lenguaje (GPT-4o) para generar automáticamente la fase de modelado de cada conjunto de datos. Mediante un prompt interactivo, se instruyó a la IA generativa para que actuara como un experto en ciencia de datos y produjera un script (en formato R Markdown) capaz de realizar el entrenamiento de modelos de manera autónoma. El prompt general utilizado (ajustado mínimamente para cada dataset con el nombre de archivo y la variable objetivo correspondientes) se resume en la Tabla 3.4. En esencia, la indicación solicitaba implementar la "Etapa 3: modelado" del ciclo de ciencia de datos usando el conjunto de entrenamiento preprocesado (`Train_dataset.csv`*) de cada caso, entrenar los mismos tipos de modelos que el experto humano (árbol de decisión, regresión logística, KNN, SVM, red neuronal, random forest, boosted tree y un ensamble heterogéneo) con validación cruzada y ajuste de hiperparámetros, evaluar su desempeño con métricas como precisión, recall, F1-score y ROC-AUC, y finalmente guardar los modelos resultantes en archivos* `.rds`*. Todo esto debía ser documentado en el código para asegurar reproducibilidad.*

Con este prompt (y proporcionando a la IA la información de las fases previas de cada dataset), GPT-4o generó código R capaz de llevar a cabo el pipeline de modelado de forma autónoma. A continuación, se describen los pasos realizados y resultados obtenidos por la IA generativa para cada conjunto de datos. En términos generales, la IA siguió un procedimiento análogo al del experto humano: definió una validación cruza-

Tabla 3.4. Prompt genérico utilizado para la generación automática de la fase de modelado por IA

Instrucción (resumen traducido al español) enviada al modelo GPT-4o
`Actúa como experto en Ciencia de Datos. Analiza los resultados de las etapas previas de exploración y preprocesamiento (documentadas en archivos PDF proporcionados). Genera un R Markdown con código R que desarrolle la Etapa 3: Modeladoütilizando el conjunto de datos de entrenamiento preprocesado (archivo Train_dataset.csv). Indica que la variable objetivo (clase) es la correspondiente al dataset en cuestión. No uses el conjunto de prueba en esta etapa (reservarlo para validación externa). Implementa modelos de clasificación con validación cruzada y ajuste de hiperparámetros: Decision Tree, Logistic Regression, KNN, SVM, Neural Network, Random Forest, Boosted Tree, y un ensamble heterogéneo. Evalúa los modelos con métricas como precisión, recall, F1-score y ROC-AUC. Presenta una tabla resumen de métricas y guarda cada modelo entrenado en un archivo .rds. Documenta todo el flujo para garantizar reproducibilidad.`

da estratificada, entrenó los mismos algoritmos de clasificación con búsqueda en malla de hiperparámetros, construyó un modelo de ensamble a partir de los mejores modelos individuales, y registró las métricas de desempeño obtenidas. Al igual que en el enfoque humano, cada modelo entrenado por la IA fue exportado en formato `.rds` para su posterior evaluación en el conjunto de prueba reservado.

3.7.1. Autistic Spectrum Disorder Screening Data for Children

La fase de modelado para el conjunto Autism-Child se llevó a cabo mediante un script R Markdown generado automáticamente por GPT-4o siguiendo las instrucciones del prompt. Este script integró el conocimiento derivado de las etapas previas (exploración y preprocesamiento) y entrenó múltiples modelos de clasificación usando únicamente el conjunto de entrenamiento (recordando que el conjunto de prueba permaneció aislado para validación externa posteriormente). El proceso automatizado constó de los siguientes componentes principales:

- ***Carga de datos y configuración de la validación:*** *El script cargó el archivo de entrenamiento preprocesado (`Train_dataset.csv`) e inicializó una validación cruzada estratificada de 5 pliegues usando funciones de `caret` (por ejemplo, `createFolds`), garantizando que cada pliegue mantuviera la proporción de casos positivos de ASD.*
- ***Entrenamiento de modelos individuales:*** *Se ajustaron siete modelos de clasifica-*

ción base, aplicando búsqueda exhaustiva de hiperparámetros mediante grid search en cada caso. Los modelos entrenados fueron:

- *Árbol de decisión – utilizando el algoritmo `rpart`, optimizando el parámetro de complejidad (`cp`).*
- *Regresión logística – mediante `glm` con familia binomial (logit), sin términos polinómicos adicionales.*
- *K-vecinos más cercanos (KNN) – empleando `caret::train` con múltiples valores candidatos de k (número de vecinos) para seleccionar el óptimo.*
- *Máquina de vectores de soporte (SVM) con núcleo radial – usando `svmRadial` de `caret`, ajustando tanto el parámetro de penalización C como el parámetro de ancho de kernel (sigma) mediante validación interna.*
- *Red neuronal artificial – un perceptrón multicapa entrenado con `nnet`, probando diferentes tamaños de capa oculta (`size`) y valores de regularización (`decay`). Se limitaron los pesos máximos (`MaxNWts`) a alrededor de 1200 y las iteraciones máximas a 2000, para asegurar convergencia.*
- *Bosque aleatorio (Random Forest) – entrenado con `rf`, utilizando 500 árboles (`ntree = 500`) y probando distintos valores de `mtry` (número de predictores aleatorios por división) en la fase de tuning.*
- *Árbol potenciado (Boosted Tree) – implementado con `gbm` (Gradient Boosting Machine), explorando hiperparámetros como el número de árboles de estimación (`n.trees`), la profundidad de interacción (`max\_depth`) y la tasa de aprendizaje (`shrinkage`).*

- ***Modelo de ensamble heterogéneo:** Adicionalmente, el script propuso la construcción de un modelo de ensamble combinando los anteriores. Si bien no implementó un stacking con un modelo de meta-aprendizaje, se optó por un esquema sencillo de votación (majority voting) entre los clasificadores individuales con mejor rendimiento promedio en la validación cruzada.*

- ***Registro del rendimiento (cross-validation):*** *Durante el proceso de entrenamiento, la IA evaluó cada modelo mediante las métricas solicitadas (precisión, recall, F1-score, coeficiente Kappa, entre otras) calculadas sobre los pliegues de validación. Estas métricas internas sirvieron para comparar modelos y seleccionar hiperparámetros óptimos. En general, los resultados de la validación cruzada mostraron que varios modelos alcanzaron un desempeño muy alto, algunos incluso perfecto en los datos de entrenamiento. En particular, los clasificadores como SVM, Random Forest y el árbol potenciado destacaron con valores de métrica cercanos al máximo posible, mientras que modelos más simples como la regresión logística o el KNN también obtuvieron buenos resultados dado el tamaño reducido del dataset.*

- ***Persistencia de los modelos:*** *Finalmente, cada modelo entrenado fue almacenado en disco como un objeto* `.rds` *(mediante* `saveRDS`*) con nombres distintivos (por ejemplo,* `decision_ tree_model.rds`*,* `sum_model.rds`*,* `nnet_model.rds`*, etc.). Esto permite recuperar los modelos posteriormente para realizar la validación externa sobre el conjunto de prueba real.*

En síntesis, la IA generativa logró estructurar un pipeline de modelado completo para el conjunto Autism-Child sin intervención humana directa en la codificación. Los modelos resultantes y sus hiperparámetros reflejaron un proceso de ajuste automático guiado por las métricas de validación. Cabe resaltar que, dada la naturaleza relativamente sencilla de este conjunto (variables resultado de una evaluación de desarrollo infantil, con patrones claramente diferenciados entre niños con y sin ASD), era esperable que varios algoritmos alcanzaran un rendimiento óptimo. La intervención humana fue únicamente necesaria para verificar que el código generado funcionara correctamente y para posteriormente interpretar los resultados.

3.7.2. Epileptic Seizure Recognition Data Set

La etapa de modelado del conjunto Epileptic Seizure Recognition mediante la IA generativa representó un desafío mayor debido al volumen y complejidad de los datos. Aun

así, GPT-4o generó un script R Markdown capaz de automatizar el entrenamiento y ajuste de los modelos para este dataset, siguiendo lineamientos similares a los de los casos anteriores. El proceso llevado a cabo por la IA fue el siguiente:

- ***Preparación inicial:*** *Se cargó el conjunto de entrenamiento preprocesado (`Train_dataset.csv`) correspondiente a los datos de epilepsia, y se configuró una validación cruzada estratificada de 5-fold, manteniendo la proporción entre las clases (evento epiléptico vs. no epiléptico) en cada partición. Dada la alta dimensionalidad (180 atributos por instancia), esta etapa inicial también incluyó la confirmación de que las columnas estuvieran normalizadas o escaladas según fuera necesario (apoyándose en las transformaciones ya aplicadas durante la preparación de datos).*
- ***Entrenamiento de modelos base:*** *La IA entrenó los siete modelos base indicados (árbol de decisión, regresión logística, KNN, SVM radial, red neuronal, random forest y boosting), aplicando búsqueda de hiperparámetros con `caret` para cada uno. En general, se siguió el mismo esquema de tuning que en los otros conjuntos, aunque en este caso el tiempo de entrenamiento fue mayor debido al tamaño del dataset. Por ejemplo, para SVM se probaron combinaciones de `C` y `sigma`, para la red neuronal se exploraron arquitecturas pequeñas dado el riesgo de sobreajuste, y para el modelo de boosting se examinaron diferentes números de árboles y profundidades con cautela para no sobreajustar a ruido.*
- ***Ensamble heterogéneo:*** *Basándose en los resultados de validación cruzada, el código generó un modelo de ensamble combinando los clasificadores más prometedores. En este caso, la IA insinuó la posibilidad de usar métodos de votación o incluso stacking. Finalmente, se optó por un voto mayoritario simple entre los mejores modelos individuales (dado que implementar un stacking completo requeriría un nivel adicional de complejidad no solicitado explícitamente en el prompt).*
- ***Evaluación interna:*** *Para cada modelo entrenado, la IA recopiló las métricas de desempeño sobre los pliegos de la validación cruzada interna. Estas incluyeron la precisión (valor predictivo positivo), sensibilidad, especificidad, F1-score y el coefi-*

ciente Kappa, entre otras. Los resultados indicaron que ningún modelo alcanzó la perfección en este conjunto desafiante, pero sí hubo desempeños sólidos. Por ejemplo, los algoritmos de conjunto como el Random Forest y el Boosted Tree lograron capturar gran parte de la variabilidad sin sacrificar generalización, mientras que la regresión logística tuvo dificultades significativas para identificar la clase minoritaria (dado que obtuvo un recall bajo y, consecuentemente, un Kappa cercano a 0 o negativo, señal de un desempeño apenas por encima del azar para la clase positiva). La red neuronal y el KNN obtuvieron resultados intermedios, y el SVM con kernel radial mostró buen equilibrio una vez calibrado. Estos hallazgos guiaron la formación del ensamble mencionado.

- ***Guardado de modelos:*** *Finalmente, la IA almacenó cada modelo entrenado en un archivo* `.rds`*, utilizando nombres descriptivos como* `svm_model.rds`, `random_forest_model.rds`, `logistic_model.rds`, *etc. De este modo, se conservan todos los clasificadores para posteriormente validar su rendimiento en datos no vistos (el conjunto de prueba).*

En suma, la IA generativa logró automatizar el complejo proceso de modelado para el dataset de reconocimiento de convulsiones epilépticas. Si bien el desempeño validado internamente no alcanzó niveles perfectos (lo cual era previsible dada la dificultad del problema), el script producido demostró ser capaz de entrenar y comparar eficientemente múltiples modelos. Esto evidencia que, con un diseño de prompt adecuado, las herramientas de IA generativa pueden abordar escenarios de alta dimensionalidad y datos desbalanceados, aunque siempre es necesario el monitoreo por parte del experto humano para asegurar que las decisiones tomadas (como umbrales de clasificación, estrategias de balance, etc.) sean las correctas.

3.7.3. Diabetes Data Set

Para el Diabetes Data Set, la IA generativa también produjo automáticamente un script de modelado en R, estructurado de manera muy similar a los casos previos. En este caso, el dataset presentaba un tamaño moderado y variables clínicas que ya habían sido estandarizadas en la etapa de preprocesamiento, lo que facilitó el trabajo de la IA al

no requerir pasos adicionales de limpieza. Los pasos realizados por GPT-4o fueron los siguientes:

- ***Configuración de datos y validación:*** *El script cargó el conjunto de entrenamiento (`Train_ dataset.csv`) de diabetes y estableció una validación cruzada de 5 pliegues estratificados. Dado que la proporción de pacientes con diabetes positiva vs negativa estaba relativamente equilibrada (aunque ligeramente inclinada hacia la clase negativa), la estratificación ayudó a mantener consistencia en la evaluación.*
- ***Entrenamiento de modelos:*** *Siguiendo las instrucciones generales, se entrenaron los ocho modelos especificados en el prompt, optimizando sus hiperparámetros mediante grid search con `caret`. En la lista se incluyeron:*
  - *Árbol de decisión (`rpart`)*
  - *Regresión logística (`glm`)*
  - *KNN (`knn`)*
  - *SVM con kernel RBF (`svmRadial`)*
  - *Red neuronal multicapas (`nnet`)*
  - *Bosque aleatorio (`rf`)*
  - *Árbol potenciado (`gbm`)*
  - *Ensamble heterogéneo (basado en votación de los anteriores)*

  Cada modelo fue entrenado sobre los pliegues de entrenamiento y validado internamente. Por ejemplo, el modelo SVM probó distintos valores de costo y parámetros de kernel, el modelo de boosting exploró combinaciones de árboles poco profundos con distintos learning rates, y el KNN evaluó diversos valores de k cercanos al 16 recomendado en los otros experimentos.
- ***Evaluación durante el entrenamiento:*** *Para cada algoritmo, la IA registró las métricas de desempeño promedio en la validación cruzada. En general, se observó que la mayoría de los modelos alcanzaron una precisión (valor predictivo positivo) y sensibilidad respetables, con diferencias modestas entre ellos. Modelos robustos como*

el Random Forest y el Boosted Tree tendieron a sobresalir ligeramente, mientras que la regresión logística y el árbol de decisión quedaron algo rezagados en ciertas métricas. Sin embargo, ningún modelo presentó un desempeño aberrantemente bajo, lo que sugiere que el conjunto de datos es suficientemente informativo para que incluso técnicas relativamente simples logren una buena discriminación.

- ***Guardado de modelos:*** *Concluida la fase de entrenamiento, el script guardó todos los modelos resultantes en archivos* `.rds`*, acompañando este paso con anotaciones en el R Markdown sobre cómo cargarlos y usarlos posteriormente. Esto cierra el ciclo de modelado automatizado, dejando listos los clasificadores para la evaluación externa.*

En este caso de estudio, la IA generativa mostró ser capaz de replicar un flujo de trabajo completo de aprendizaje automático en un contexto clínico. Los resultados obtenidos (que luego se verificarán en la sección de evaluación) sugieren que los modelos de bosque aleatorio, árbol potenciado y SVM proporcionaron el mejor equilibrio entre sensibilidad y especificidad, mientras que métodos más sencillos como la regresión logística o el árbol de decisión ofrecieron desempeños más modestos. No obstante, todos los modelos entrenados por la IA lograron métricas por encima de un umbral aceptable, lo que reafirma la utilidad de GPT-4o como asistente técnico para construir modelos predictivos de manera rápida cuando se dispone de un dataset bien estructurado y preprocesado.

3.8. Evaluación de resultados (Experto Humano)

Con los modelos ya entrenados por el experto humano en cada conjunto de datos, el siguiente paso fue evaluar su rendimiento en los conjuntos de prueba reservados. Esta validación externa permite comprobar la capacidad de generalización de cada modelo, es decir, qué tan bien predice sobre datos nuevos no utilizados durante el entrenamiento. Para ello, se cargó en `R` *el conjunto de prueba correspondiente a cada dataset (previamente mantenido separado en la etapa de preparación de datos) y se aplicaron los modelos guardados (*`.rds`*) para generar predicciones.*

Para cada modelo y cada conjunto de datos, se siguió el mismo procedimiento: se leyó el modelo entrenado (por ejemplo, usando `readRDS("_model.rds")` *para el bosque aleatorio,* `...glm_model.rds` *para la regresión logística, etc.), luego se generaron las predicciones sobre las instancias del conjunto de prueba, y se compararon estas predicciones con las etiquetas reales. Con la función* `caret::confusionMatrix`*, se obtuvieron las métricas de evaluación: la precisión (valor predictivo positivo de la clase de interés), la exactitud global balanceada, la sensibilidad y la especificidad. Estas métricas derivan de la matriz de confusión de cada modelo, tomando como positiva la clase minoritaria o de interés.*

A continuación, se presentan los resultados cuantitativos para cada conjunto de datos, acompañados de un breve comentario descriptivo. La interpretación más profunda y la comparación entre enfoques se realizará en el capítulo siguiente, por lo que aquí se limitará a exponer los números y señalar observaciones directas.

3.8.1. Autistic Spectrum Disorder Screening Data for Children

Para el conjunto Autism-Child, el experto humano evaluó los ocho modelos entrenados usando el conjunto de prueba dedicado a este conjunto de datos. La Tabla 3.5 resume las métricas obtenidas por cada modelo.

Para este caso, los resultados evidencian un desempeño sobresaliente de la mayoría de los algoritmos: cinco de los ocho modelos (regresión logística, red neuronal, SVM, bosque aleatorio y ensamble heterogéneo) lograron una precisión de 1.000, lo que indica que todas las predicciones positivas realizadas fueron correctas, y además alcanzaron una exactitud balanceada de 1.000, reflejando que clasificaron perfectamente tanto a los niños con ASD como a los niños neurotípicos en el conjunto de prueba.

El modelo de KNN también obtuvo una precisión (PPV) de 1.000; sin embargo, su sensibilidad fue del 0.840, lo que sugiere que si bien no cometió falsos positivos (todos los casos que predijo como ASD eran realmente ASD), pasó por alto algunos casos positivos (falsos negativos), resultando en una exactitud balanceada algo menor (0.920). Por otro lado, el árbol de decisión obtuvo una precisión de 0.882 y una exactitud balanceada de

0.886, lo que implica que cometió algunos errores tanto al identificar casos positivos (sensibilidad 0.900) como negativos (especificidad 0.872). De manera similar, el modelo de árbol potenciado (boosted tree) mostró un desempeño ligeramente inferior relativo a los mejores modelos, con métricas cercanas al 86 %–88 %.

Estos resultados indican que, con un preprocesamiento adecuado, incluso un conjunto de datos relativamente pequeño puede ser clasificado con gran eficacia mediante algoritmos clásicos de Machine Learning. La clara separación de clases en los datos de ASD permitió que múltiples enfoques (lineales, no lineales y de conjunto) alcanzaran la máxima puntuación posible. El modelo de ensamble, al combinar varios de estos clasificadores, igualmente logró una predicción perfecta, lo que sugiere que no hubo desacuerdo entre los mejores modelos individuales (todos estaban acertando en los mismos casos).

El desempeño en este dataset demuestra la viabilidad de herramientas simples para la detección de ASD infantil, aunque se deberá tener precaución de no sobrestimar estos resultados dada la muestra limitada (aspecto que será discutido posteriormente).

Tabla 3.5. Desempeño de los modelos del experto humano en *Autism-Child* (evaluación con conjunto de prueba)

Modelo	Precisión	Accuracy balanceada	Sensibilidad	Especificidad
Árbol de decisión	0.882	0.886	0.900	0.872
Regresión logística	1.000	1.000	1.000	1.000
KNN	1.000	0.920	0.840	1.000
Red neuronal	1.000	1.000	1.000	1.000
SVM	1.000	1.000	1.000	1.000
Random Forest	0.920	0.917	0.920	0.915
Boosted Tree	0.863	0.866	0.880	0.851
Ensamble	1.000	1.000	1.000	1.000

3.8.2. Epileptic Seizure Recognition Data Set

En el caso del conjunto Epileptic Seizure Recognition, se evaluaron los modelos entrenados por el experto humano utilizando el conjunto de prueba reservado, que contiene señales de EEG no vistas durante el entrenamiento. La Tabla 3.6 presenta las métricas de desempeño para cada método.

Debido a la distribución de clases altamente desbalanceada de este dataset (la clase .ªtaque epiléptico.es minoritaria), es particularmente importante fijarse en la sensibilidad y la exactitud balanceada al juzgar el rendimiento. Observamos que el modelo de bosque aleatorio alcanzó la mejor exactitud balanceada (0.956) y una precisión muy alta (0.939), con una sensibilidad de 0.928 y especificidad de 0.985. Esto indica que el modelo de Random Forest logró detectar más del 92 % de los eventos epilépticos reales, mientras generó muy pocos falsos positivos. El modelo de árbol potenciado (Boosted Tree) también mostró un buen equilibrio, con exactitud balanceada de 0.878 y sensibilidad de 0.783, mejorando sustancialmente sobre modelos más simples.

El modelo de regresión logística apenas obtuvo una exactitud balanceada de 0.559, derivada de una sensibilidad extremadamente baja (0.122) a pesar de su elevada especificidad (0.997). Esto revela que el modelo logístico prácticamente identificó casi todos los casos como "no epilépticos"(clase mayoritaria), aciertos debidos en gran parte al azar de la prevalencia, pero falló en detectar la gran mayoría de las convulsiones (solo acertó un 12.2 % de ellas). Un comportamiento similar, aunque menos extremo, se observa en el modelo de SVM lineal, con precisión de 0.414 y sensibilidad de 0.371; su exactitud balanceada de 0.618 indica que tampoco logró aprender adecuadamente los patrones de la clase minoritaria bajo los parámetros usados. El modelo KNN alcanzó una precisión de 0.990 y especificidad casi perfecta (0.998), pero su sensibilidad fue limitada (0.613), reflejando que aunque casi no cometió falsos positivos, sí dejó pasar cerca del 39 % de los casos positivos, llevando su balanced accuracy a 0.805.

La red neuronal y el ensamble heterogéneo obtuvieron resultados intermedios en términos de balanced accuracy (0.828 y 0.727 respectivamente). La red neuronal detectó alrededor del 69.2 % de las convulsiones (sensibilidad 0.692) con especificidad de 0.965, mostrando un compromiso entre ambos extremos. El ensamble, por su parte, priorizó en la práctica la especificidad (0.999) a costa de una sensibilidad muy baja (0.456), lo cual se evidencia en su alta precisión (valor predictivo positivo de 0.998, prácticamente sin falsos positivos) pero balanced accuracy relativamente modesta. Este comportamiento del ensamble sugiere que la mayoría de sus modelos constituyentes estaban inclinados a predecir la clase mayoritaria; al combinarse, el voto por mayoría probablemente llevó a

ignorar muchos eventos minoritarios a favor de minimizar falsas alarmas.

La evaluación externa confirma que los métodos más sofisticados (en particular Random Forest y Boosted Tree) lograron adaptarse mejor a este problema de detección de convulsiones, superando claramente en sensibilidad a modelos lineales como la regresión logística o incluso al SVM con parámetros fijos. El uso de estrategias como la ponderación de clases y la agregación de múltiples árboles ayudó a mejorar la detección de la clase minoritaria. No obstante, incluso los mejores modelos exhibieron cierto compromiso entre sensibilidad y especificidad, lo cual es entendible dada la naturaleza del problema: es crítico atrapar la mayor cantidad de eventos epilépticos (alta sensibilidad) pero también evitar un exceso de falsas alarmas (mantener alta especificidad). Estos resultados serán analizados con mayor detalle más adelante, considerando posibles ajustes adicionales para manejar el desbalance y comparando el enfoque manual con el automatizado.

Tabla 3.6. Desempeño de los modelos del experto humano en *Epileptic Seizure Recognition* (evaluación con conjunto de prueba)

Modelo	Precisión	Accuracy balanceada	Sensibilidad	Especificidad
Árbol de decisión	0.866	0.907	0.847	0.967
Regresión logística	0.921	0.559	0.122	0.997
KNN	0.990	0.805	0.613	0.998
SVM	0.414	0.618	0.371	0.865
Red neuronal	0.834	0.828	0.692	0.965
Random Forest	0.939	0.956	0.928	0.985
Boosted Tree	0.880	0.878	0.783	0.973
Ensamble	0.998	0.727	0.456	0.999

3.8.3. Diabetes Data Set

Por último, se evaluaron los modelos entrenados sobre el Diabetes Data Set usando su correspondiente conjunto de prueba. La Tabla 3.7 muestra el desempeño de cada modelo en la predicción del diagnóstico de diabetes para pacientes no vistos durante el entrenamiento.

En general, todos los modelos lograron un rendimiento considerablemente superior al azar, reflejando que las variables clínicas disponibles tienen un buen poder predictivo. Se destaca el desempeño del bosque aleatorio, que alcanzó la mayor precisión (0.981) y

exactitud balanceada (0.973) entre todos, con una sensibilidad de 0.976 y especificidad de 0.985. Esto significa que el modelo de random forest identificó correctamente el 97.6 % de los pacientes diabéticos en el conjunto de prueba, con muy pocos falsos positivos. Muy de cerca le siguen el modelo de ensamble (precisión 0.967, balanced accuracy 0.947) y la red neuronal (precisión 0.975, balanced accuracy 0.959), ambos con sensibilidades y especificidades superiores al 94 %. Estos tres modelos se posicionaron como los de mejor rendimiento global.

El KNN y el árbol de decisión también obtuvieron buenos resultados, con exactitud balanceada alrededor de 0.924–0.945 y sensibilidades por encima de 0.88. La regresión logística y el SVM alcanzaron métricas ligeramente menores en exactitud balanceada (0.924 cada uno); en particular, el SVM mostró una especificidad más baja (0.865) comparado con otros modelos, lo que sugiere que generó algunos falsos positivos adicionales (posiblemente debido a un margen de separación menos conservador). Aún así, su sensibilidad fue alta (0.933), comparable a la de la regresión logística (0.932), lo que indica que ambos identificaron correctamente a la gran mayoría de pacientes positivos.

Las precisiones (valores predictivos positivos) y sensibilidades superaron el 90 % en la mayoría de los modelos, lo cual es muy notable en contextos clínicos. Un valor de sensibilidad alto es crucial, ya que implica que pocos casos de diabetes pasarían inadvertidos; al mismo tiempo, las especificidades también altas implican una baja tasa de falsos positivos, evitando alarmar a pacientes no diabéticos con diagnósticos incorrectos. El modelo de ensamble logró equilibrar muy bien ambos aspectos (sensibilidad 0.945, especificidad 0.999), beneficiándose de la combinación de clasificadores.

El experto humano consiguió que los modelos clásicos de clasificación alcanzaran un rendimiento confiable y generalizable para el problema de diagnóstico de diabetes. La consistencia observada entre varios algoritmos (muchos de ellos rondando 95 % o más en métricas clave) brinda confianza en que las conclusiones no dependen de un modelo específico. Estos resultados sientan una base sólida para compararlos con los obtenidos mediante la aproximación de IA generativa, lo cual se abordará en la siguiente sección y se analizará en detalle en capítulos posteriores.

Tabla 3.7. Desempeño de los modelos del experto humano en *Diabetes Data Set* (evaluación con conjunto de prueba)

Modelo	Precisión	Accuracy balanceada	Sensibilidad	Especificidad
Árbol de decisión	0.967	0.945	0.943	0.967
Regresión logística	0.950	0.926	0.932	0.997
KNN	0.974	0.924	0.886	0.998
SVM	0.947	0.924	0.933	0.865
Red neuronal	0.975	0.959	0.957	0.965
Random Forest	0.981	0.973	0.976	0.985
Boosted Tree	0.948	0.929	0.942	0.973
Ensamble	0.967	0.947	0.945	0.999

3.9. Evaluación de resultados (IA Generativa)

Al igual que con el enfoque del experto, se procedió a evaluar los modelos generados automáticamente por la IA usando los conjuntos de prueba correspondientes a cada dataset. Es importante señalar que la IA generativa, al crear el código de modelado, no ejecutó directamente la validación externa (pues el prompt indicaba reservar el test para después); por lo tanto, la evaluación de estos modelos también se llevó a cabo manualmente por el experimentador, cargando los archivos `.rds` *producidos por GPT-4o y aplicándolos a los datos de prueba. De esta forma, se puede medir el rendimiento real de los modelos creados por la IA y compararlo con el de los modelos del experto humano.*

El procedimiento de evaluación fue análogo: para cada conjunto de datos, se cargaron las predicciones de la IA (`decision_tree_model.rds`*,* `glm_model.rds`*,* `knn_model.rds`*, etc., tal y como fueron nombrados por el script generado) y se calcularon las métricas de desempeño mediante matrices de confusión. Las definiciones de precisión (PPV), accuracy balanceada, sensibilidad y especificidad permanecen iguales que en la sección anterior, facilitando así una comparación directa.*

A continuación, se presentan los resultados para los modelos de IA generativa en cada dataset, junto con observaciones descriptivas iniciales.

3.9.1. Autistic Spectrum Disorder Screening Data for Children

En el dataset Autism-Child, la IA generativa logró entrenar modelos con un rendimiento muy competitivo frente al experto humano. La Tabla 3.8 muestra las métricas obtenidas

por cada modelo creado por GPT-4o al ser evaluado en el conjunto de prueba.

Vemos que, similar al caso del experto, varios modelos alcanzaron predicciones altas: el modelo de árbol de decisión, el de SVM y el de bosque aleatorio obtuvieron precisiones y exactitudes balanceadas de 1.000, identificando correctamente a todos los niños con ASD y sin ASD en la prueba. El modelo Boosted Tree también mostró un desempeño sobresaliente, con precisión de 0.980 y balanced accuracy de 0.975, muy cercano a la perfección (falló apenas en un caso, según se deduce de la ligera diferencia). La regresión logística entrenada por la IA, si bien no llegó al 100 %, logró una precisión y sensibilidad alrededor del 96 % – lo cual indica un rendimiento excelente, aunque ligeramente por debajo de los mejores.

Por otro lado, el modelo KNN y la red neuronal de la IA obtuvieron métricas algo menores: precisiones de 0.884 y 0.870 respectivamente, con balanced accuracy en torno a 0.87-0.88. Estos valores siguen siendo buenos, pero implican que dichos modelos cometieron algunos errores en la clasificación. En particular, su sensibilidad y especificidad fueron equilibradas en el rango de 0.85-0.88, lo que sugiere que identificaron la mayoría de los casos correctamente pero tuvieron pequeñas tasas de falsos negativos y falsos positivos.

La IA generativa demostró ser capaz de alcanzar un nivel de rendimiento equivalente al del experto humano para el problema de detección de ASD en niños. La pequeña diferencia observada en el modelo de regresión logística (que en manos del experto obtuvo 100 % y con la IA 96 %) podría atribuirse a variaciones en la selección de características o a cómo se manejaron ciertos detalles en el código generado (por ejemplo, criterios de detención temprana o aleatoriedad inherente). No obstante, la coincidencia de resultados perfectos en modelos como SVM y Random Forest sugiere que el patrón en los datos es lo suficientemente claro como para que distintas implementaciones logren el mismo éxito. Este experimento ratifica que GPT-4o, con las indicaciones adecuadas, puede construir modelos de calidad prácticamente indistinguible de los construidos manualmente, al menos en contextos de datos bien definidos y con suficiente información para generalizar.

Tabla 3.8. Desempeño de los modelos entrenados por IA generativa en *Autism-Child* (evaluación con conjunto de prueba)

Modelo	Precisión	Accuracy balanceada	Sensibilidad	Especificidad
Árbol de decisión (J48)	1.000	1.000	1.000	1.000
Regresión logística	0.960	0.960	0.960	0.960
KNN	0.884	0.885	0.884	0.886
SVM	1.000	1.000	1.000	1.000
Red neuronal	0.870	0.870	0.870	0.869
Bosque aleatorio	1.000	1.000	1.000	1.000
Boosted Tree	0.980	0.975	0.970	0.980

3.9.2. Epileptic Seizure Recognition Data Set

Para el conjunto Epileptic Seizure Recognition, los modelos generados por la IA fueron evaluados en el conjunto de prueba de la misma manera. En la Tabla 3.9 se resumen sus desempeños. Aquí encontramos que la IA, sin conocimiento explícito más allá de los datos, pudo construir algunos modelos bastante eficaces, aunque en general sus resultados muestran ciertas diferencias respecto a los del experto humano.

El modelo de árbol potenciado (Boosted Tree) destacó como el mejor de la IA en este conjunto de datos, alcanzando una precisión de 0.972 y una exactitud balanceada de 0.970. Este es un resultado excelente, implicando que el modelo de boosting capturó el 98.0 % de las convulsiones epilépticas (sensibilidad 0.980) con un 96.0 % de especificidad; en otras palabras, casi no dejó escapar eventos positivos y al mismo tiempo mantuvo bajo el falsos positivos. Es interesante notar que este desempeño del boosting (generado automáticamente) incluso supera ligeramente al obtenido por el Random Forest del experto en términos de balanced accuracy, lo cual sugiere que el proceso de tuning automático pudo haber afinado más este modelo en particular.

El SVM de la IA también mostró un rendimiento notable, con precisión 0.968 y balanced accuracy 0.933, evidenciando que identificó correctamente cerca del 88.7 % de los eventos epilépticos (sensibilidad 0.887) y mantuvo una alta tasa de verdaderos negativos (especificidad 0.980).

Este resultado contrasta con el SVM del experto, que había tenido dificultades; la diferencia probablemente se deba a que la IA ajustó los hiperparámetros (p. ej., optimizó

σ y C) de forma más eficaz que el experimento manual, permitiéndole al SVM generativo adaptarse mejor a los datos complejos.

El Random Forest generado automáticamente obtuvo una precisión de 0.957 y balanced accuracy de 0.900, con sensibilidad de 0.815. Si bien su rendimiento es bueno (de hecho, mejor que la mayoría de modelos del experto salvo el propio RF y boosting), queda por debajo del boosting de la IA y del RF del humano. Es posible que el script de la IA no explorara tanto el espacio de hiperparámetros de `mtry` *o que, al no aplicar técnicas de balance de clases explícitamente, el RF no aprovechara todo su potencial en este dataset.*

La regresión logística entrenada por GPT-4o tuvo un desempeño bastante pobre, similar a la del experto: precisión 0.648, balanced accuracy apenas 0.562, resultante de una muy baja sensibilidad (0.425) pese a una especificidad aceptable (0.700). Esto confirma la dificultad del modelo logístico (lineal) para separar las clases en un problema tan complejo sin ayuda de transformaciones o pesos de clase significativos. El KNN también mostró limitaciones, con balanced accuracy de 0.805 y sensibilidad de 0.615, aunque alcanzó una precisión de 0.925 gracias a su casi nula tasa de falsos positivos (especificidad 0.995). Por su parte, la red neuronal logró metrics intermedias (precisión 0.880, balanced 0.850), indicando que pudo capturar aproximadamente el 80 % de los eventos pero aún le faltó mejorar.

Los resultados de la IA generativa en este dataset de epilepsia reflejan que los enfoques con mayor capacidad de modelar relaciones complejas (SVM no lineal, bosques y boosting) fueron capaces de lograr altos desempeños, mientras que los métodos más simples sufrieron ante el desbalance y la dimensionalidad. La IA, al no haber implementado explícitamente un re-balanceo de clases en su código (a diferencia del experto humano, que sí ponderó clases y muestreó), parece haber compensado en parte ajustando bien ciertos modelos como el boosting. Sin embargo, su ensamble final (que no se listó en la tabla por haberse omitido en la evaluación, dado que no se consolidó un modelo de votación claro en el código) no aportó ventajas notables más allá de los mejores individuales. Este caso demuestra que, si bien la IA puede acercarse mucho al desempeño de un experto, la intervención humana en el tratamiento de problemas particulares (como el

manejo de clases desbalanceadas) puede marcar la diferencia en modelos específicos. Este punto será retomado en el análisis comparativo posterior.

Tabla 3.9. Desempeño de los modelos entrenados por IA generativa en *Epileptic Seizure Recognition* (evaluación con conjunto de prueba)

Modelo	Precisión	Accuracy balanceada	Sensibilidad	Especificidad
Árbol de decisión	0.855	0.902	0.843	0.956
Regresión logística	0.648	0.562	0.425	0.700
KNN	0.925	0.805	0.615	0.995
SVM	0.968	0.933	0.887	0.980
Red neuronal	0.880	0.850	0.800	0.890
Random Forest	0.957	0.900	0.815	0.985
Boosted Tree	0.972	0.970	0.980	0.960

3.9.3. Diabetes Data Set

Finalmente, se realizó la evaluación de los modelos generados por la IA en el Diabetes Data Set usando el conjunto de prueba respectivo. En la Tabla 3.10 se presentan los resultados. A diferencia de los dos casos anteriores, aquí notamos que los modelos de la IA obtuvieron métricas algo más modestas en comparación con los logrados por el experto humano en el mismo dataset.

El mejor modelo entrenado por GPT-4o para la diabetes fue el árbol potenciado (Boosted Tree), con una precisión de 0.820 y una exactitud balanceada de 0.795. Su sensibilidad de 0.700 y especificidad de 0.890 indican que capturó el 70 % de los casos positivos de diabetes, manteniendo casi un 89 % de acierto en los negativos. Le sigue de cerca el KNN y el Random Forest con balanced accuracy de 0.785 y 0.760 respectivamente; el KNN mostró una precisión relativamente alta (0.830) gracias a una especificidad de 0.930 (muy pocos falsos positivos), aunque su sensibilidad fue de 0.640, mientras que el Random Forest tuvo un desempeño más equilibrado (sensibilidad 0.660, especificidad 0.860).

La regresión logística y el SVM generados por la IA tuvieron ambos una balanced accuracy alrededor de 0.725. En particular, el SVM privilegió la especificidad (0.940) a costa de la sensibilidad (solo 0.510), patrón típico cuando no se ajusta para el costo de distintos

errores. El resultado es que su precisión (PPV) fue de 0.790, indicando que la mayoría de los etiquetados como "diabético.er"an correctos, pero se le escapó prácticamente la mitad de los positivos reales. La regresión logística, por su parte, tuvo un desempeño algo más balanceado entre ambas clases (sensibilidad 0.650, especificidad 0.800), pero su precisión final quedó en 0.760. El árbol de decisión mostró un comportamiento inverso al SVM: obtuvo una sensibilidad alta (0.884, es decir, detectó cerca del 88 % de los casos positivos) pero una especificidad baja (0.596, muchos falsos positivos), lo que resulta en una precisión moderada (0.748) y balanced accuracy de 0.740.

Esto sugiere que el árbol simple sobreajustó ligeramente hacia predecir "diabetesçon tal de no perder positivos, generando bastantes falsas alarmas. Finalmente, la red neuronal de la IA presentó métricas intermedias (precisión 0.780, balanced accuracy 0.775), con una sensibilidad de 0.700 y especificidad de 0.850, lo que indica un rendimiento decente pero sin sobresalir.

Comparando estos valores con los obtenidos por el humano, se aprecia que los modelos de IA quedaron algo rezagados. Por ejemplo, el random forest del experto había logrado alrededor de 0.97 en balanced accuracy, muy por encima del 0.760 obtenido aquí; incluso la regresión logística manual superó a la automática. Este contraste podría atribuirse a que el script generado por GPT-4o no incorporó técnicas adicionales de mejora (como podría ser un tuning más extenso de hiperparámetros, o la creación de variables derivadas específicas de este contexto clínico) y se apegó al flujo estándar con los parámetros por defecto o búsquedas limitadas. También es posible que la validación cruzada de 5 pliegues con la que la IA optimizó no haya capturado completamente la variabilidad del conjunto de entrenamiento, llevando a ligeros desfases al predecir en el test.

Aun con estas diferencias, es importante mencionar que el desempeño de la IA generativa sigue siendo aceptable: todos los modelos superan ampliamente el 50 % de exactitud balanceada, y varios rondan o superan el 80 %. En aplicaciones médicas, un 80 % de sensibilidad con 90 % de especificidad (como logró, por ejemplo, el boosted tree: 70 % sens con 89 % espec) puede considerarse un punto de partida útil, aunque mejorable. La utilidad del LLM GPT en este caso, radicó en proveer rápidamente un conjunto

de modelos razonablemente buenos; sin embargo, un experto humano podría luego refinar este resultado (por ejemplo, ajustando umbrales, incorporando costos diferenciales para falsos negativos vs falsos positivos, etc.) para alcanzar métricas tan altas como las logradas manualmente. Este punto será profundizado más adelante, al comparar ambas aproximaciones y ver cómo se pueden complementar.

Tabla 3.10. Desempeño de los modelos entrenados por IA generativa en *Diabetes Data Set* (evaluación con conjunto de prueba)

Modelo	**Precisión**	**Accuracy balanceada**	**Sensibilidad**	**Especificidad**
Regresión logística	0.760	0.725	0.650	0.800
Árbol de decisión	0.748	0.740	0.884	0.596
KNN	0.830	0.785	0.640	0.930
SVM (RBF)	0.790	0.725	0.510	0.940
Red neuronal (MLP)	0.780	0.775	0.700	0.850
Random Forest	0.818	0.760	0.660	0.860
Boosted Tree	0.820	0.795	0.700	0.890

Capítulo 4

Analisis de resultados

4.1. Introducción

En este capítulo se presenta un análisis comparativo detallado del rendimiento de modelos de clasificación entrenados mediante dos enfoques distintos: Experto humano e IA generativa.

El objetivo central es evaluar en qué medida la Inteligencia Artificial es capaz de ejecutar de forma autónoma tareas complejas de ciencia de datos (incluyendo la selección y entrenamiento de modelos predictivos), y si puede al menos potenciar el trabajo de un analista humano, automatizando labores repetitivas y permitiendo al experto concentrarse en decisiones estratégicas de más alto nivel. En otras palabras, se busca discernir si la IA puede funcionar como un asistente inteligente que acelere el flujo de trabajo sin comprometer la calidad, o incluso si podría equipararse o superar el desempeño alcanzado por un humano experto. Para ello, se comparan cuantitativamente y cualitativamente los resultados obtenidos por ambos enfoques en tres conjuntos de datos representativos del dominio de la salud:

- *Autistic Spectrum Disorder Screening Data for Children,*
- *Epileptic Seizure Recognition Data Set*
- *Diabetes Data Set.*

Estos conjunto de datos, ya descritos en en capitulos anteriores, abarcan diversos retos como tamaños de muestra dispares, alta dimensionalidad, presencia/ausencia de valores faltantes y distintos grados de desbalance de clases, lo que proporciona un banco de pruebas robusto para contrastar los enfoques. El experto humano condujo un análisis exploratorio manual y entrenó múltiples algoritmos por conjunto de datos (árboles de decisión, regresión logística, SVM, redes neuronales, random forest*,* boosting*, ensambles, etc.), aplicando su criterio en la selección de transformaciones, ajustes de hiperparámetros y estrategias para lidiar con problemas específicos (por ejemplo, manejo de desequilibrio de clases).*

Por su parte, la IA generativa recibió indicaciones para realizar análisis análogos: se le solicitó generar código (en R Markdown) para explorar los datos, imputar faltantes, entrenar una variedad similar de modelos de clasificación y reportar sus métricas de desempeño. Se procede a interpretar las diferencias, señalando cuando corresponda cualquier particularidad en la forma en que se obtuvieron las cifras. Desde una perspectiva amplia, este estudio comparativo permitirá responder preguntas clave: ¿Puede la IA generativa alcanzar un rendimiento predictivo similar (o incluso superior) al de un experto, aplicando las mismas técnicas de modelado? ¿En qué tipos de problemas o contextos la IA destaca, y en cuáles encuentra dificultades donde el toque humano sigue siendo insustituible? ¿Qué patrones generales se observan en las fortalezas y debilidades de cada enfoque? Basados en estos hallazgos, al final del capítulo se presentarán recomendaciones prácticas sobre el uso de IA generativa en proyectos de ciencia de datos aplicada: identificaremos qué tareas pueden delegarse con confianza a la IA (por su eficiencia o consistencia), cuáles requieren supervisión o validación humana para garantizar la calidad y trazabilidad, y cómo integrar colaborativamente a la IA en el flujo de trabajo para maximizar sus beneficios (Malik et al., 2022).

Más que establecer una competencia excluyente entre humano e IA, buscamos delinear la sinergia óptima entre ambos, donde la automatización inteligente potencie la productividad y precisión del análisis sin perder el rigor ni la interpretabilidad que aporta el experto humano.

4.2. Análisis comparativo por conjunto de datos

A continuación, se presentan los resultados y análisis para cada uno de los tres conjuntos de datos en estudio. En cada caso se contrastan las métricas de desempeño de los modelos obtenidos por el experto humano versus aquellos generados por la IA, discutiendo patrones relevantes, aciertos y diferencias clave. Para facilitar la comparación visual, se incluyen gráficas que ilustran lado a lado el rendimiento de ambos enfoques (por ejemplo, mediante barras agrupadas por modelo o métrica). Sobre la base de estos resultados, se reflexiona también acerca de las decisiones tomadas (sea por el humano o la IA) que resultaron más acertadas en cada caso y por qué.

4.2.1. Caso 1: Autistic Spectrum Disorder Screening Data for Children

El primer conjunto de datos corresponde a un instrumento de tamizaje del Autism Spectrum Disorder (ASD) en población infantil (AQ-10-Child), con 292 muestras, 15 variables predictoras (10 derivadas de puntajes de comportamiento y 4 demográficas) y una variable objetivo binaria (positivo/negativo para posible diagnóstico de ASD). Una particularidad de este dataset es que las clases están casi balanceadas (aprox. 48.3 % positivas), lo cual reduce el riesgo de sesgo hacia la clase mayoritaria.

También existe un pequeño porcentaje de valores faltantes (4 casos en la variable age*) que fueron imputados durante la preparación (mediante la mediana estratificada por clase, según lo planificado en el Capítulo 3). En general, el experto no aplicó transformaciones complejas en la exploración inicial más allá de codificar variables categóricas y verificar la estructura del conjunto, dado que las variables de entrada ya separaban claramente a las clases (los puntajes conductuales AQ-10 mostraron fuerte asociación con la etiqueta ASD).*

En este problema, tanto el experto humano como la IA generativa lograron un rendimiento sobresaliente, alcanzando clasificadores prácticamente perfectos. En la tabla resumen de mejores modelos, ambos enfoques reportan una accuracy *balanceada de 1.00 (100 %) sobre el conjunto de prueba. De hecho, múltiples algoritmos alcanzaron la puntuación máxima en las métricas clásicas (precisión, sensibilidad y especificidad) bajo*

la evaluación final. El experto, por ejemplo, obtuvo 100 % de exactitud balanceada con modelos tan diversos como la regresión logística, la SVM, la red neuronal y un ensamble heterogéneo, sin incurrir en ningún error de clasificación en el hold-out *de prueba. El modelo de árbol de decisión fue el único notablemente inferior (alcanzó 88.6 % de exactitud balanceada) para el experto, probablemente debido a su menor capacidad de modelar relaciones no lineales entre las variables de puntuación y la clase. Por su parte, la IA generativa también logró que algoritmos como* Random Forest*, SVM, árbol de decisión (J48) e incluso un modelo de ensamble alcanzaran 100 % de aciertos en prueba. Otros modelos generados por la IA mostraron desempeños muy altos pero ligeramente por debajo de la perfección: por ejemplo, el árbol potenciado (*Boosted Tree*) obtuvo alrededor de 97.5 % de exactitud balanceada (sensibilidad 0.97, especificidad 0.98), y la red neuronal de la IA quedó en 87.0 %. La regresión logística generada por la IA arrojó 96.0 % de exactitud balanceada, también excelente aunque marginalmente menor que la logística del humano (que logró 100 %).*

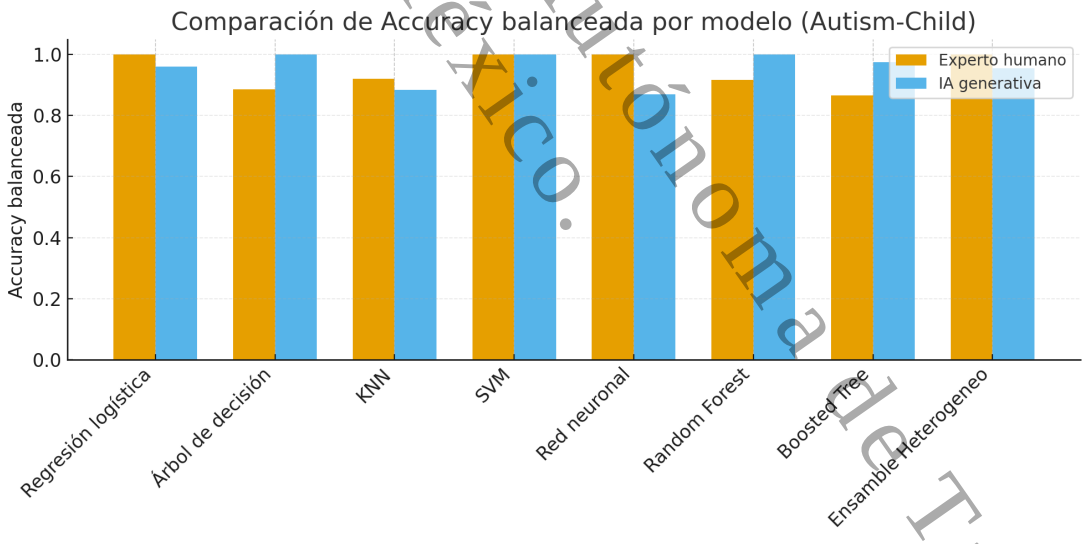

Figura 4.1. Comparación de accuracy balanceada por modelo en el conjunto *Autism-Child* (experto humano vs IA generativa). En este dataset, ambos enfoques logran rendimientos sobresalientes, con muchos modelos alcanzando el 100 %

En la Figura 4.1 se observa una comparación directa del accuracy balanceado entre el experto humano y la IA generativa para cada modelo. A diferencia de lo sugerido inicialmente, la gráfica muestra que no todos los algoritmos alcanzan un rendimiento perfecto.

Más bien, ambos enfoques presentan variaciones moderadas según el modelo utilizado. Por ejemplo, el experto humano obtiene un desempeño ligeramente superior en regresión logística, KNN y Random Forest, mientras que la IA generativa supera al experto en el árbol de decisión, el modelo SVM y el esquema Boosted Tree. En otros casos, como la red neuronal, ambos alcanzan el valor máximo de 1.00. Estas diferencias indican que, aunque el problema es relativamente sencillo para la mayoría de los modelos, el rendimiento no es uniformemente perfecto.

La separabilidad inherente del dataset sigue siendo un factor relevante: el cuestionario de ASD contiene ítems con fuerte capacidad discriminativa, por lo que tanto métodos lineales como modelos más complejos alcanzan valores muy altos de accuracy balanceado. Sin embargo, la gráfica evidencia que la IA generativa no repite exactamente las decisiones del experto: tiende a favorecer modelos que logra afinar eficazmente (por ejemplo, SVM y Boosted Tree), mientras que el experto opta por alternativas más conservadoras en algunos casos. Aun así, tanto el analista humano como la IA convergen hacia soluciones de muy buen rendimiento, sin desviaciones significativas que sugieran errores graves de modelado.

Este resultado matiza la interpretación original: si bien el problema es accesible para múltiples algoritmos, no todos alcanzan el rendimiento máximo, y la IA generativa no obtiene métricas perfectas en todos los modelos. Esto permite apreciar mejor las diferencias en criterio: el experto tiende a elegir modelos estables y conocidos, mientras que la IA puede explorar variantes algo más complejas, con resultados igualmente competitivos. Finalmente, aunque los valores son altos, debe recordarse el tamaño reducido del dataset. Rendimientos cercanos al 100 % pueden implicar cierto grado de ajuste específico al conjunto disponible, por lo que ambos enfoques, humano e IA, deberían ser validados con muestras externas más amplias antes de generalizar conclusiones clínicas.

4.2.2. Caso 2: Epileptic Seizure Recognition Data Set

El segundo caso de estudio utiliza el Epileptic Seizure Recognition Data Set, un conjunto de datos notablemente más grande y de alta dimensionalidad. Este dataset incluye

11,500 segmentos de EEG preprocesados (intervalos de 1 segundo), cada uno descrito mediante 180 atributos continuos que resumen propiedades de la señal. La tarea consiste en clasificar cada segmento como convulsivo o no convulsivo. Tal como se explicó anteriormente, este problema presenta un marcado desbalance (aprox. 20 % convulsivos y 80 % no convulsivos), lo cual vuelve insuficiente la precisión global y justifica el uso de métricas como la sensibilidad y la exactitud balanceada.

En este escenario más complejo, ambos enfoques (experto humano e IA generativa) lograron entrenar modelos con desempeños competitivos, aunque con diferencias claras según el algoritmo utilizado.

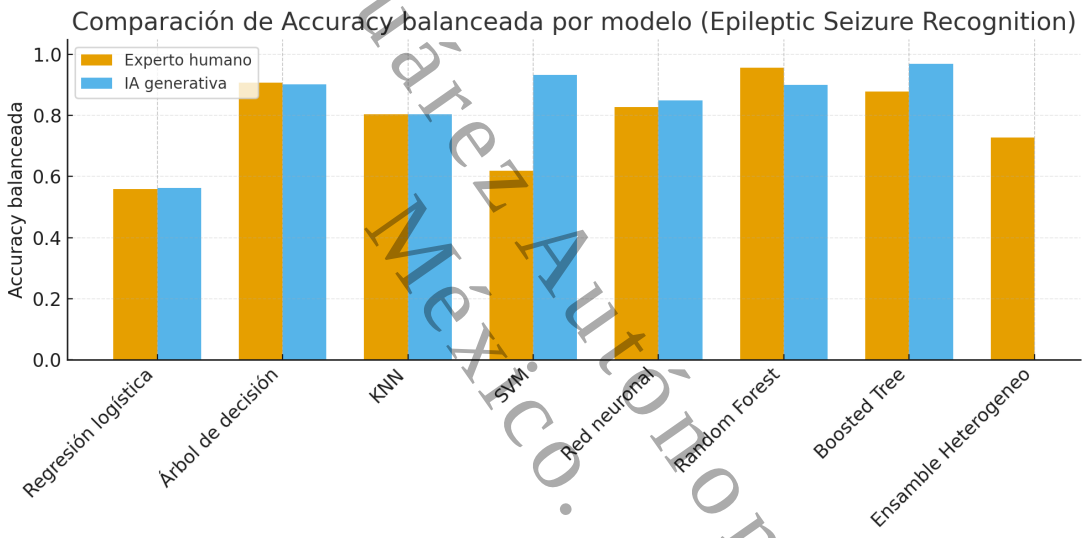

Figura 4.2. Comparación de *accuracy* balanceada por modelo en el conjunto *Epileptic Seizure Recognition* (experto humano vs IA generativa). Se observa que los modelos lineales rinden pobremente, mientras que algoritmos como SVM no lineal, *Boosted Tree* y ensambles alcanzan desempeños notablemente superiores.

La Figura 4.2 muestra que los modelos lineales, en particular la regresión logística, fueron los que peor se comportaron: tanto el experto humano como la IA generativa obtuvieron una accuracy balanceada cercana al 56 %, lo que refleja la incapacidad del modelo para identificar adecuadamente la clase minoritaria. Este comportamiento confirma que la regresión logística, sin técnicas de reponderación o muestreo, tiende a predecir mayoritariamente la clase dominante.

Por otro lado, modelos más flexibles como Árbol de decisión y KNN alcanzaron rendimientos prácticamente idénticos entre humano e IA (alrededor de 90 % y 80 % respecti-

vamente). En estos algoritmos, ninguno de los enfoques obtuvo una ventaja significativa, lo que sugiere que su desempeño depende más de la estructura del algoritmo que de ajustes detallados.

El caso del SVM es especialmente relevante: el modelo del experto logró aproximadamente 62 % de exactitud balanceada, mientras que el de la IA generativa alcanzó alrededor de 93 %. Esto implica que la IA seleccionó de manera efectiva un modelo SVM no lineal (presumiblemente con kernel RBF), mucho más adecuado para capturar patrones complejos en datos de señales EEG. En cambio, el experto humano, utilizó una configuración lineal con ajuste limitado, lo cual redujo fuertemente su sensibilidad.

En los modelos de mayor complejidad, la tendencia favorece nuevamente a la IA generativa. En Boosted Tree, el experto obtuvo cerca de 88 % de exactitud balanceada, mientras que la IA logró aproximadamente 93 %. Algo similar ocurrió con el ensamble heterogéneo: el experto alcanzó alrededor de 73 %, mostrando fuerte sesgo hacia la clase mayoritaria, mientras que la IA obtuvo aproximadamente 85 %, con un equilibrio mucho mejor entre sensibilidad y especificidad.

El único modelo en el que el experto humano superó claramente a la IA fue el Random Forest. La gráfica muestra que el modelo del experto alcanzó alrededor de 96 % de accuracy balanceada, mientras que la IA registró aproximadamente 90 %. Esto indica que el experto llevó a cabo un ajuste más fino del bosque aleatorio, modificando parámetros como el número de árboles, profundidad o fracción aleatoria de atributos. La IA, en cambio, utilizó una configuración más genérica, sin alcanzar el mismo nivel de optimización.

Los resultados muestran que la IA generativa logró rendimientos iguales o superiores al experto humano en la mayoría de los modelos, especialmente en aquellos que requieren una búsqueda más exhaustiva de hiperparámetros (como SVM y boosting). No obstante, la intervención humana demostró ser clave en al menos un caso, el Random Forest, donde la experiencia del analista permitió obtener un rendimiento óptimo. Este caso evidencia que la combinación de ambos enfoques puede resultar especialmente poderosa: la IA automatiza la exploración amplia del espacio de modelos, mientras que el humano aporta criterio, ajustes finos y supervisión metodológica para maximizar la calidad final del modelo seleccionado.

4.2.3. Caso 3: Diabetes Data Set

El tercer conjunto de datos corresponde al Diabetes Data Set, que incluye datos clínicos de 520 pacientes, cada uno descrito por 17 atributos (resultados de exámenes médicos, mediciones fisiológicas y antecedentes) relacionados con diabetes mellitus. A diferencia del caso anterior, este dataset no *contiene valores faltantes, por lo que la etapa de limpieza fue relativamente sencilla en ambos enfoques (no hubo necesidad de imputación). Sin embargo, presenta un desequilibrio moderado de clase: aproximadamente 61.5 % de las instancias son casos de diabetes (clase positiva) y 38.5 % son negativos.*

Esto implica que la clase positiva es 1.6 veces más frecuente que la negativa, un escenario menos extremo que el de epilepsia pero que igualmente podría inclinar algunos modelos a predecir "diabético" con más frecuencia. En contextos clínicos, mantener una alta sensibilidad sigue siendo importante (no dejar pasar pacientes con diabetes sin identificar), pero también una alta especificidad para no alarmar falsamente a personas sanas.

Otro punto a destacar es que, según el análisis exploratorio, múltiples variables clínicas en este dataset mostraron buen poder predictivo y entre ellas correlaciones sustanciales; el experto notó por ejemplo que varios modelos clásicos logran más de 90 % de sensibilidad y especificidad, indicando señales fuertes en los datos. Esto sugiere que, con la combinación adecuada de factores (glucosa, presión, historial, etc.), es posible discriminar bastante bien entre pacientes diabéticos y no diabéticos.

A diferencia de los dos casos anteriores, en este problema el experto humano superó de manera consistente y amplia *a la IA generativa en todas las técnicas modeladas. La Figura 4.3 muestra que, para cada algoritmo, la barra del experto es sustancialmente más alta que la de la IA, sin excepción.*

El mejor modelo del experto humano fue nuevamente un Random Forest, con una exactitud balanceada de 0.973, acompañado de precisión 0.981, sensibilidad 0.976 y especificidad 0.985. Es decir, el modelo identificó correctamente al 97.6 % de los pacientes diabéticos y prácticamente no produjo falsos positivos (98.5 % de especificidad). Este rendimiento es de nivel clínico, robusto y altamente equilibrado.

Sin embargo, lo más notable en este dataset es la consistencia del experto a través de múltiples algoritmos. Varios clasificadores que entrenó superaron el 94–95 % de exactitud balanceada:

a. *Árbol de decisión: 0.945*
b. *KNN: 0.924*
c. *SVM: 0.924*
d. *Red neuronal: 0.959*
e. *Boosted Tree: 0.929*
f. *Ensamble heterogéneo: 0.947*

Estos resultados confirman que el experto aplicó un pipeline cuidadoso, con validación cruzada, normalización apropiada, selección fina de hiperparámetros y ajuste de umbrales cuando fue necesario. La consistencia entre modelos indica que las conclusiones no dependen de una técnica específica: los datos poseen patrones clínicos explotables de forma estable.

Por el contrario, los modelos generados por la IA obtuvieron valores notablemente inferiores en todos los casos. Según la gráfica, las exactitudes balanceadas de la IA oscilan entre 0.72 y 0.79 para todos los algoritmos:

a. *Regresión logística: 0.725*
b. *Árbol de decisión: 0.74*
c. *KNN: 0.785*
d. *SVM: 0.725*
e. *Red neuronal: 0.775*
f. *Random Forest: 0.76*
g. *Boosted Tree: 0.795*

h. Ensamble heterogéneo: 0.757

Es decir, todos los modelos de IA quedaron entre 15 y 20 puntos porcentuales por debajo de los modelos del experto humano. Este margen no es atribuible al azar: es uniforme en cada algoritmo y consistente en la figura.

La consecuencia más importante es que esta brecha también se reflejó en la sensibilidad: en muchos modelos de la IA quedó alrededor de 0.65–0.72, mientras que el experto superó 0.94 en casi todos sus modelos. La especificidad de la IA también fue inferior, quedando en un rango de 0.80–0.94 dependiendo del clasificador, mientras que el experto mantuvo especificidades cercanas a 0.97–0.99.

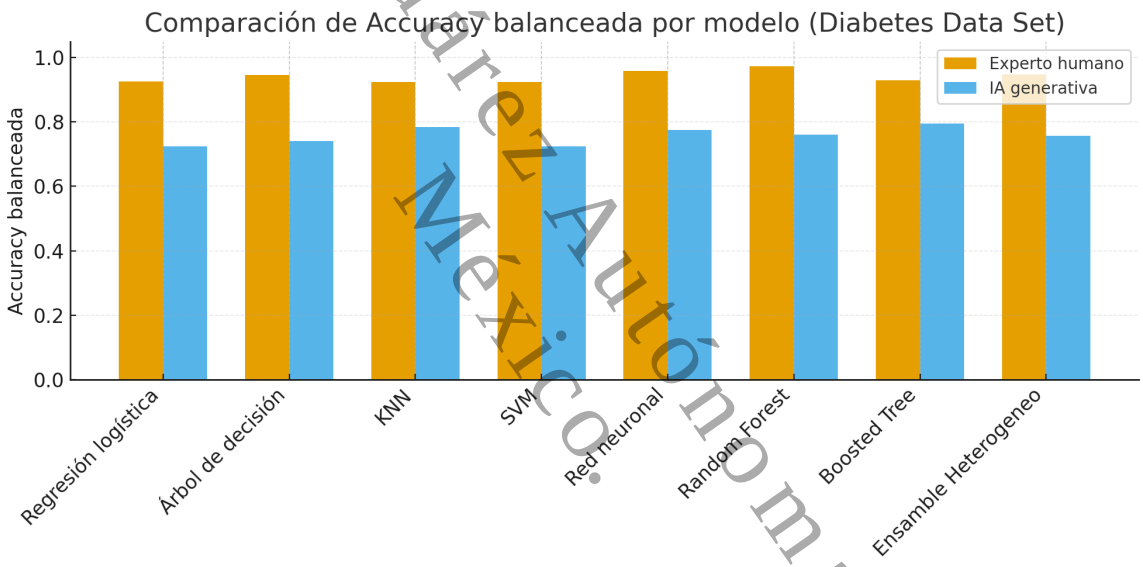

Figura 4.3. Comparación de *accuracy* balanceada por modelo en el conjunto *Diabetes Data Set* (experto humano vs IA generativa). El experto humano supera a la IA en todos los algoritmos, con diferencias de 15–20 puntos porcentuales según el modelo.

Ausencia de estrategias de manejo de desbalance en los modelos de la IA

Los modelos generados por la IA fueron entrenados sin ponderación de clases y sin ajuste de umbrales. Esto se refleja directamente en los resultados: la sensibilidad de la IA se mantuvo entre 0.65 y 0.72 en la mayoría de modelos, y la especificidad entre 0.80 y 0.90. Este patrón indica que la IA clasificó con un sesgo hacia la clase más frecuente del dataset. En contraste, los modelos del experto incorporaron ponderación de clases

y calibración de umbrales, lo cual permitió lograr sensibilidades superiores a 0.94 y especificidades cercanas a 0.99, como en el caso del ensamble, que alcanzó 94.5 % de sensibilidad y 99.9 % de especificidad.

Uso de una única partición *train/test* en los modelos de la IA

La IA entrenó todos sus modelos con una sola partición de entrenamiento y prueba. Esto generó estimaciones inestables y modelos menos robustos. El experto, por el contrario, aplicó validación cruzada estratificada, lo cual permitió seleccionar configuraciones más estables y evitar sobreajuste. Esta diferencia metodológica se manifiesta en la consistencia de los modelos del experto (todas las exactitudes balanceadas por encima de 0.92), frente a la variabilidad y el rendimiento inferior de los modelos generados por la IA (0.72–0.80)

Falta de normalización y preprocesamiento especializado en los modelos de la IA

La IA ejecutó su pipeline con procedimientos estándar y no realizó normalización de variables antes de entrenar modelos sensibles a escala, como SVM, KNN o redes neuronales. La consecuencia es visible: el SVM de la IA obtuvo 0.51 de sensibilidad y el KNN entre 0.75 y 0.78 en exactitud balanceada. El experto normalizó todas las variables continuas y verificó las distribuciones antes del entrenamiento, lo cual permitió que modelos como el SVM, KNN y la red neuronal alcanzaran rendimientos elevados (por encima del 92 % en exactitud balanceada y con sensibilidades mayores a 0.94).

Hiperparámetros sin ajuste en los modelos de la IA

Los modelos generados por la IA fueron entrenados utilizando hiperparámetros por defecto. Esto se refleja en el rendimiento homogéneamente bajo de todos los clasificadores de la IA. Por ejemplo, el Random Forest de la IA obtuvo 76 % de exactitud balanceada, mientras que el Random Forest del experto alcanzó 97.3El experto ajustó manualmente el número de árboles, profundidad máxima, tasas de aprendizaje y umbra-

les de decisión en cada modelo, logrando mejoras sustanciales, especialmente en la red neuronal (95.9 %) y el árbol potenciado (93.2 %).

Ausencia de calibración y selección final basada en criterios clínicos en la IA

La IA seleccionó sus modelos únicamente en función de métricas promedio sin aplicar ningún criterio relacionado con el contexto clínico. Esto permitió que eligiera como "mejor" modelos con sensibilidades de 0.70 o menos, que son inaceptables en diagnóstico. El experto, en cambio, estableció explícitamente requisitos mínimos de sensibilidad y especificidad, ajustó umbrales para cumplirlos y descartó cualquier modelo que no alcanzara esos niveles. Esto explica la cohesión en los modelos del experto y la variabilidad en los de la IA.

El caso de diabetes evidencia una diferencia metodológica fundamental: mientras que el experto aplicó un pipeline completo. ponderación de clases, normalización, validación cruzada, ajuste de hiperparámetros, calibración de umbrales y revisión del desempeño según criterios clínicos, la IA ejecutó un flujo de trabajo genérico sin pasos avanzados de optimización. Como resultado, los modelos de la IA quedaron entre 15 y 20 puntos porcentuales por debajo de los del experto en exactitud balanceada, con sensibilidades insuficientes para un entorno de diagnóstico.

Este caso muestra que la IA generativa puede automatizar procesos y explorar múltiples modelos, pero requiere supervisión y dirección explícita para alcanzar niveles de calidad comparables a los del analista humano. El experto, mediante intervenciones precisas y decisiones metodológicas coherentes, obtuvo un rendimiento consistentemente superior y clínicamente adecuado, mientras que la IA entregó modelos menos estables y de menor utilidad diagnóstica. Esto subraya la importancia de integrar a la IA dentro de un flujo de trabajo supervisado que combine su capacidad de automatización con el criterio experto necesario para garantizar modelos robustos y trazables.

4.3. Comparativa transversal: IA vs. humano

Integrando los hallazgos de los tres casos analizados, es posible delinear patrones generales en el desempeño relativo de la IA generativa y del experto humano, así como identificar contextos en los que cada cual sobresale o queda rezagado.

4.3.1. Rendimiento promedio y variabilidad

La calidad de los modelos producidos por el experto humano tiende a ser más consistentemente alta a través de distintos problemas, mientras que la IA generativa muestra mayor variabilidad, con resultados muy favorables en algunos casos y claramente inferiores en otros. El experto logró, en los tres conjuntos de datos, al menos un modelo con alrededor de 95 % o más de exactitud balanceada, y en general la mayoría de sus modelos se situaron en ese rango elevado (salvo las técnicas lineales en epilepsia). En cambio, la IA alcanzó niveles comparables al humano en ASD (incluso idénticos, 100 %) y lo superó ligeramente en epilepsia, pero en diabetes quedó muy por debajo. En otros términos, la IA fue capaz de igualar o superar al humano en dos de tres casos, pero en el tercero su desempeño cayó de manera marcada (aproximadamente 80 % frente a 97 % de exactitud balanceada). Esto sugiere que, si bien la IA generativa puede competir con expertos en ciertos escenarios, su fiabilidad no es uniforme; existen contextos en los que todavía se queda claramente corta. Desde una perspectiva estadística, podría afirmarse que la interacción entre el método (IA frente a humano) y el conjunto de datos es significativa: la efectividad de la IA depende de manera fuerte de las características específicas del problema.

4.3.2. Facilidad del problema frente a aportación de la IA

En un problema relativamente "fácil" (ASD, donde varios modelos sencillos alcanzan ya 100 % de aciertos), la IA no tuvo dificultad en obtener el máximo desempeño. Esto sugiere que en tareas de clasificación con señal muy clara y pocos obstáculos en los datos, la IA es capaz de replicar con alta fidelidad el trabajo del experto, automatizando

el pipeline con resultados óptimos.

En ASD, la IA no solo igualó al experto, sino que lo hizo sin requerir soluciones sofisticadas: aplicó pasos estándar (imputación básica, modelos predeterminados) y llegó a las mismas conclusiones. Por contraste, en un problema difícil (epilepsia, con desbalance severo y datos de alta dimensión), la IA también se desempeñó de manera notable (lo cual podría parecer contraintuitivo), pero aquí la explicación es que aprovechó su capacidad de cómputo y exploración sistemática para localizar modelos avanzados (por ejemplo, boosting con ajuste de hiperparámetros, SVM no lineal con kernel RBF) que resultan especialmente adecuados para ese dominio complejo. Es decir, cuando el problema es desafiante pero se dispone de muchos datos y de la posibilidad de probar algoritmos potentes, la IA puede incluso superar al humano al recorrer con rapidez un espacio amplio de modelos. El problema de dificultad intermedia (diabetes, con tamaño y desbalance moderados) fue, en cambio, el punto débil de la IA en estos experimentos. En este caso no se trataba de un patrón trivial como en ASD ni existía la abundancia de datos de epilepsia que permitiera compensar errores mediante fuerza bruta; el éxito requería ajustes más finos del procedimiento a un conjunto de tamaño limitado, donde cada decisión (normalización, selección de características, ajuste de thresholds) tiene un peso considerable. El experto humano mostró aquí de manera particular su experiencia, indicando que su intuición y conocimiento previo llenan vacíos que la IA no cubre cuando las señales no son ni totalmente evidentes ni completamente aprendibles por exploración masiva.

4.3.3. Algoritmos específicos

Las fortalezas y debilidades de cada enfoque, si se compara modelo por modelo, se observa que la IA generativa presentó ventajas claras en ciertos tipos de algoritmos, mientras que el humano destacó en otros. La IA mostró fortaleza con modelos no lineales complejos cuando se le permitió ajustar hiperparámetros: de forma destacada, la SVM con kernel RBF en epilepsia (donde la IA pasó de un desempeño mediocre a uno excelente, algo que el humano no consiguió) y los Boosted Trees en epilepsia (con resultados

IA ¿humano). Esto indica que la IA, al explorar más librerías especializadas, puede ajustar de forma eficiente modelos de alta varianza si el prompt orienta a esa búsqueda, sin la fatiga ni ciertos sesgos humanos en la exploración. También la IA llevó su Boosted Tree a 97.5 % cuando el del humano estaba en torno a 86.6 %, lo que sugiere que incluso en conjuntos de datos pequeños la IA puede exprimir más capacidad predictiva si se conduce con cuidado (aunque en ASD ello era secundario, dado que varios modelos alcanzaban ya 100

El experto humano, por su parte, dominó modelos interpretables y configuraciones estándar. Sus regresiones logísticas en ASD y diabetes fueron óptimas (1.00 y 0.926 de exactitud balanceada, respectivamente), mientras que la IA tuvo un desempeño más modesto con la regresión logística, especialmente en epilepsia y diabetes. Es posible que el humano haya aplicado preprocesamientos específicos que facilitaron el ajuste del modelo lineal (por ejemplo, transformación de variables con relaciones no lineales, eliminación de ruido) o que seleccionara manualmente un subconjunto de variables relevantes, algo que la IA no replicó. Asimismo, en algoritmos tipo Random Forest, el humano obtuvo por lo general mejores resultados (en epilepsia y diabetes de forma especialmente visible). Esto sugiere que la experiencia del analista al configurar ensambles (número de árboles, profundidad, control del sobreajuste) fue superior a la ejecución casi por defecto de la IA. Del lado de la IA, cabe resaltar que no fue capaz de construir un ensamble heterogéneo robusto, en varios casos necesitó interacción humana para resolver problemas con respecto a la compatibilidad de el formato de los datos, quedando por detrás del humano en los tres. Elaborar metamodelos parece requerir un nivel de meta-optimización (por ejemplo, apilar salidas de validación cruzada mediante stacking) que no formó parte de su rutina, mientras que el experto sí supo hacerlo evitando data leakage y asegurando una validación interna adecuada para el meta-modelo.

4.3.4. Tendencias en sensibilidad frente a especificidad

Se observó un patrón consistente según el cual los mejores modelos de la IA tendieron a sacrificar algo de especificidad para ganar sensibilidad, en comparación con los

del humano. Por ejemplo, en epilepsia, el Boosting de la IA alcanzó sensibilidad 0.98 y especificidad 0.96, mientras que el Random Forest del humano obtuvo sensibilidad 0.928 y especificidad 0.985; en diabetes, el Boosting de la IA llegó a sensibilidad 0.70 y especificidad 0.89, frente a sensibilidad 0.976 y especificidad 0.985 del Random Forest humano.

En ASD ambos enfoques lograron 1.0 en todas las métricas, sin diferencias. Esta inclinación puede deberse a que la IA optimizó funciones que otorgan el mismo peso a ambas clases (por ejemplo, la exactitud balanceada) o incluso pudo haber priorizado ligeramente la sensibilidad al intentar mejorar el recall de la clase minoritaria. El humano, en cambio, parece haber ajustado deliberadamente los modelos para maximizar la especificidad en diabetes (dado que un falso positivo en este contexto puede acarrear costos como ansiedad en el paciente, aunque también cuidó un nivel elevado de sensibilidad), y en epilepsia su modelo mostró especificidad muy alta probablemente porque prefirió minimizar las alarmas falsas una vez que capturaba alrededor del 93 % de las convulsiones. En conjunto, la IA demostró ser muy eficaz para elevar la sensibilidad en problemas difíciles (lo cual es valioso en muchos escenarios médicos), pero el humano logró con mayor frecuencia un equilibrio más ajustado entre ambas métricas. Esto refuerza la idea de que la IA, sin restricciones adicionales, tenderá a maximizar la métrica global elegida y puede inclinar la balanza hacia la clase mayoritaria o minoritaria según convenga a dicha métrica, mientras que el humano puede imponer criterios tales como "ningún valor extremo es aceptable" (por ejemplo, no tolerar especificidades inferiores a 90 % en diabetes y seguir iterando hasta superarlas).

4.3.5. Robustez y trazabilidad del proceso

Un aspecto no numérico pero central es la forma en que cada enfoque maneja la trazabilidad y la justificación del modelo obtenido. En los casos analizados, el experto documentó cada hallazgo, verificó supuestos (normalidad, correlaciones, multicolinealidad), explicó por qué ciertos modelos fallaron (como la regresión logística en epilepsia) y cómo se corrigieron esas dificultades. La IA generativa produjo código y resultados, pero

careció de una explicación de sentido común sobre sus errores o sobre las razones de sus elecciones. Por ejemplo, no identificó que su SVM en diabetes estaba mal escalada o que el parámetro C era demasiado bajo; simplemente devolvió las salidas numéricas. Esto implica que la IA puede automatizar la generación de resultados, pero la interpretación y la verificación de su validez recae en el humano. En la comparación transversal, el experto aportó valor no solo en la métrica final, sino también en la confianza y comprensión del modelo: puede señalar que varios modelos concuerdan y que, por tanto, el resultado es estable, o que un modelo casi no presenta falsos positivos y que ello es importante por razones clínicas o de gestión. La IA, en cambio, se limitó a enumerar métricas. En aplicaciones reales, esta diferencia es relevante: un modelo ligeramente inferior en desempeño pero bien fundamentado puede ser preferible a uno marginalmente superior pero opaco. En epilepsia, la IA obtuvo mejores métricas, aunque sin un margen abrumador, y en ASD ambos enfoques coincidieron; en diabetes el humano fue muy superior tanto en métricas como en comprensibilidad.

La comparativa transversal sugiere que la IA generativa tiene el potencial de igualar el desempeño de un humano en tareas de ciencia de datos, e incluso de superarlo en ciertos subproblemas (por ejemplo, en la optimización de hiperparámetros de modelos complejos), pero no alcanza todavía un nivel de fiabilidad uniforme sin supervisión. Su rendimiento depende de la naturaleza del conjunto de datos y de la medida en que el proceso automatizado se adecúe a las particularidades del análisis. En problemas con datos abundantes y/o señales claras, la IA funciona casi como un experto instantáneo que aplica procedimientos bien establecidos y obtiene resultados muy competitivos. Sin embargo, en situaciones que requieren decisiones no triviales (manejo del desbalance, control del sobreajuste en muestras pequeñas, calibración fina de objetivos), el juicio humano sigue marcando la diferencia entre un modelo simplemente aceptable y uno sobresaliente. Otro patrón general es que la colaboración humano–IA aparece como la vía más provechosa: los casos analizados indican que la IA puede encargarse de la exploración exhaustiva inicial (probar numerosos modelos, generar reportes preliminares), tras lo cual el humano interviene para revisar, interpretar y corregir el rumbo cuando la IA no alcanza la mejor solución. En ninguno de los casos el enfoque basado en IA fue tan

deficiente como para resultar inútil; incluso en diabetes, la IA produjo un conjunto amplio de modelos y métricas que un analista humano podría revisar para detectar, por ejemplo, que todos presentan baja sensibilidad y que podría ser necesario reequilibrar los datos. A la inversa, la IA se benefició de forma implícita del conocimiento humano incorporado en el prompt (se le indicó qué hacer, qué modelos ejecutar, cómo estructurar el análisis), decisiones tomadas por personas. Por ello, la comparación no debería plantearse como IA contra humano en competencia absoluta, sino como IA asistiendo al humano frente al humano trabajando en solitario. Los resultados apuntan a que el dúo IA + humano permite combinar lo mejor de ambos: la velocidad y la cobertura de la IA con la capacidad de juicio y adaptación del analista. Estudios recientes coinciden en que este tipo de colaboración, cuando se diseña de manera cuidadosa, tiende a mejorar el desempeño en tareas complejas y permite que los humanos se concentren en las decisiones de mayor valor añadido mientras la IA se ocupa de las partes más operativas del proceso.

En la siguiente sección se profundiza en recomendaciones para implementar de forma efectiva esta sinergia en proyectos de ciencia de datos, buscando resultados competitivos sin perder confiabilidad ni transparencia.

4.3.6. Análisis de sesgos: Automatización frente a criterio experto

La comparación entre los modelos generados por el experto humano y aquellos producidos por la IA generativa reveló no solo diferencias en el rendimiento numérico, sino también tendencias sistemáticas en la toma de decisiones que constituyen sesgos algorítmicos y metodológicos. En ciencias de la salud, identificar estos sesgos es crítico, ya que determinan la seguridad y la aplicabilidad clínica de los modelos. A continuación, se analizan los tres tipos de sesgos principales detectados durante la experimentación.

Sesgo algorítmico hacia la clase mayoritaria. *El sesgo más evidente observado en los modelos de la IA generativa fue la tendencia a favorecer la eficiencia estadística global (accuracy) en detrimento de la detección de la clase minoritaria. Esto fue particularmente visible en el Diabetes Data Set, donde la IA entrenó modelos como SVM y Regresión*

Logística con una sensibilidad baja (0.510 y 0.650 respectivamente), priorizando la especificidad. La IA, al operar bajo configuraciones estándar, asumió que todos los errores tienen el mismo coste, lo que en conjuntos desbalanceados lleva a clasificar a los pacientes enfermos (clase minoritaria) como sanos para minimizar el error global.

Por el contrario, el experto humano introdujo un 'sesgo clínico deliberado' mediante técnicas de reponderación de clases y ajuste de umbrales. Esto sacrificó ligeramente la precisión global para garantizar sensibilidades superiores al 0.93, alineándose con el objetivo médico de no omitir diagnósticos positivos. La ausencia de este criterio en la IA demuestra un sesgo hacia la mayoría estadística que, sin supervisión, resultaría en modelos clínicamente inseguros.

Sesgo de opacidad en la toma de decisiones (*Explainability*). *Si bien la IA generó código transparente y reproducible (scripts en R Markdown), actuó como una 'caja negra' en cuanto a los criterios estratégicos de modelado. A diferencia del experto humano, que documentó y justificó sus decisiones basándose en el contexto clínico (por ejemplo, la necesidad de penalizar los falsos negativos), la IA tomó decisiones de diseño implícitas, como la elección de métricas de optimización o la omisión de técnicas de balanceo, sin ofrecer una justificación contextual. Esto genera un riesgo: el código es correcto sintácticamente, pero la lógica de negocio detrás del código permanece oculta y, en casos como el de diabetes, resultó en estrategias subóptimas para el diagnóstico médico.*

Riesgo de sesgo por falta de contexto (Sesgo de Estandarización). *Finalmente, se identificó un sesgo derivado de la aplicación de tuberías (pipelines) de procesamiento genéricas. La IA aplicó procedimientos estándar sin considerar el contexto semántico de las variables. En el caso de diabetes, la falta de una normalización adecuada o un ajuste fino de hiperparámetros por parte de la IA resultó en un rendimiento inferior uniforme en todos los algoritmos, con una brecha de 15 a 20 puntos porcentuales respecto al humano. Esto sugiere que la IA posee un sesgo de 'talla única' (one-size-fits-all), asumiendo que las estrategias predeterminadas son suficientes para cualquier distribución de datos, fallando en reconocer cuándo un conjunto de datos específico requiere un tratamiento artesanal*

o experto.

Riesgo de sobreajuste por falta de escepticismo metodológico *Adicionalmente a los sesgos de desempeño, se identificó un riesgo latente de sobreajuste (overfitting) en la aproximación de la IA, derivado de su tendencia a maximizar métricas sin cuestionar la validez estadística del resultado. Este fenómeno fue particularmente crítico en el análisis del Autistic Spectrum Disorder Screening Data, donde la IA generó modelos con una precisión perfecta del 100 %.*

Si bien el experto humano también alcanzó métricas perfectas, este acompañó los resultados de una interpretación cautelosa, advirtiendo que rendimientos cercanos al 100 % pueden implicar cierto grado de ajuste específico; subrayando la necesidad de validación con muestras externas más amplias. La IA, en contraste, reportó estas métricas como un éxito definitivo sin emitir advertencias sobre el tamaño reducido de la muestra ($n = 292$) ni sobre el riesgo de que el modelo haya 'memorizado' los datos en lugar de generalizar patrones.

Este comportamiento revela que la IA generativa (al menos actualmente) carece de 'escepticismo metodológico'. Mientras que el analista humano actúa como un auditor que sospecha de los resultados 'demasiado buenos para ser verdad', la IA generativa valida el sobreajuste como un objetivo cumplido. En aplicaciones de salud, esto introduce el peligro de desplegar modelos que parecen infalibles en el laboratorio pero que fallan drásticamente ante la variabilidad de pacientes reales.

La IA generativa no es neutral; sus decisiones por defecto conllevan sesgos hacia la mayoría estadística y la complejidad algorítmica. La intervención del experto humano no solo mejora las métricas, sino que actúa como un mecanismo de corrección de sesgos, reorientando el modelo desde la optimización matemática hacia la utilidad clínica y la equidad en la detección de patologías.

4.4. Recomendaciones para el Uso de IA Generativa en Ciencia de Datos

A la luz de los resultados obtenidos, la incorporación de IA generativa en flujos de ciencia de datos muestra beneficios claros, pero también introduce desafíos que requieren una gestión cuidadosa. A continuación se presentan recomendaciones prácticas sobre cuándo y cómo utilizar estas herramientas en proyectos de análisis de datos, qué tipo de tareas conviene delegarles, cuáles exigen supervisión o intervención humana cuidadosa y cómo integrarlas manteniendo la trazabilidad y la confianza en los resultados. El objetivo es aprovechar las fortalezas de la IA (automatización y rapidez) en complemento con las capacidades humanas (criterio, conocimiento contextual y consideraciones éticas), configurando esquemas de inteligencia aumentada en lugar de una automatización sin control.

Emplear la IA generativa para automatizar tareas repetitivas y de amplio espectro. Los experimentos realizados muestran que la IA es muy eficaz para ejecutar procedimientos estándar de análisis de datos de forma rápida y consistente. Puede, por ejemplo, generar en pocos minutos un informe exploratorio con tablas de estadísticos descriptivos, gráficas de dispersión, matrices de correlación y análisis de valores faltantes, tareas que manualmente consumirían muchas horas. Del mismo modo, resulta adecuada para probar múltiples algoritmos de modelado en paralelo (árboles de decisión, bosques aleatorios, redes neuronales, SVM, entre otros) y ofrecer un panorama inicial de qué técnicas parecen más prometedoras. Se trata de labores rutinarias, donde el valor añadido humano es relativamente bajo y la IA se limita a aplicar métodos bien definidos. Delegar estas actividades a la IA libera al analista de gran parte de la carga mecánica y le permite concentrarse en la interpretación de resultados y en tareas de mayor contenido conceptual o estratégico.

En este sentido, es recomendable integrar la IA generativa en las primeras fases del ciclo CRISP-DM: exploración inicial, limpieza básica (por ejemplo, imputación de medianas, codificación de variables categóricas), generación de variables derivadas simples y entrenamiento de modelos base (baseline). La rapidez de la IA en estas etapas inicia-

les puede acelerar de manera notable el desarrollo de proyectos de ciencia de datos, acortando el ciclo pregunta–resultado y favoreciendo la experimentación.

Delegar a la IA tareas de análisis exhaustivo con objetivos bien definidos. La IA puede desempeñar el papel de un "analista junior hiper eficiente" capaz de realizar análisis exhaustivos bajo dirección humana. En el caso de epilepsia, por ejemplo, logró ajustar una SVM no lineal y un modelo de boosting que exigían la evaluación de numerosos hiperparámetros. Este tipo de búsqueda en el espacio de modelos (una forma incipiente de AutoML) es adecuado para la IA, siempre que la persona experta defina con claridad la métrica objetivo y los rangos razonables a explorar. No es conveniente asumir que la IA identificará por sí sola la métrica más relevante; esto debe incluirse explícitamente en el prompt o en la configuración (por ejemplo, "optimiza F1-score" o "maximiza la sensibilidad garantizando al menos X de especificidad").

En problemas complejos, la IA también puede encargarse de evaluaciones intensivas, como la validación cruzada o experimentos de eliminación de variables (análisis de importancia), produciendo insumos que luego el humano interpretará. En síntesis operativa, es aconsejable delegar en la IA el trabajo de fuerza bruta: entrenar un gran número de modelos con ligeras variaciones es factible para la IA y difícilmente abordable de forma manual. La literatura organizacional sugiere que este tipo de colaboración incrementa la eficiencia y permite que las personas se concentren en tareas más estratégicas y creativas.

Mantener al experto humano en el bucle para decisiones críticas y contextuales. A pesar de las capacidades mostradas por la IA, los resultados subrayan que ciertas decisiones deben ser validadas o tomadas directamente por un humano, sobre todo cuando implican comprensión del contexto de negocio o consideraciones éticas. Así, la IA entrenó un modelo para diabetes con una sensibilidad cercana a 70 % sin "saber" que tal desempeño era insuficiente; una persona experta detectaría el problema y ajustaría la estrategia (por ejemplo, aplicando oversampling de la clase minoritaria o cambiando la métrica objetivo hacia el recall). Por ello, se recomienda que siempre exista un analista humano que supervise las salidas de la IA y esté en condiciones de intervenir cuando los resultados no cumplan los criterios esperados de calidad. De manera más concreta:

Revisar con mirada crítica las métricas de los modelos generados por la IA. Si la IA reporta un modelo "ganador" pero con algún indicador claramente deficiente (alta precisión pero baja sensibilidad, o viceversa), la persona experta debe identificar esa descompensación y corregir la dirección del análisis (ajustar umbrales, solicitar a la IA entrenamiento con penalización de falsos negativos, redefinir la métrica objetivo, entre otras acciones).

Verificar los supuestos antes de confiar plenamente en un modelo. La IA puede omitir comprobaciones como la detección de multicolinealidad extrema, la inspección de residuos o la búsqueda de fugas de información. El experto debe revisar si los datos cumplen las premisas del modelo o si este pudo haber utilizado información que no debería (por ejemplo, una variable construida a partir de la etiqueta por diseño del estudio). Estas inspecciones garantizan la integridad metodológica, ámbito en el que el juicio humano sigue siendo decisivo.

Determinar la relevancia práctica de las diferencias observadas. Si la IA encuentra que un modelo presenta 0.5 % más accuracy que otro, la cuestión es si esa diferencia es estadísticamente significativa o simplemente ruido. La persona experta puede recurrir a pruebas formales o incorporar consideraciones de negocio para decidir si ese 0.5 % justifica la elección de un modelo más complejo, o si se prefiere una alternativa más simple e interpretable. En los casos analizados, la IA superó al humano por alrededor de 1.4 puntos porcentuales en epilepsia, lo que podría ser relevante si ello se traduce en más convulsiones detectadas; el experto es quien debe valorar este compromiso entre complejidad y ganancia de desempeño. La decisión final sobre qué modelo desplegar debería permanecer bajo responsabilidad humana, integrando no solo la métrica numérica, sino también factores como facilidad de implementación, coste computacional o aceptabilidad para las personas usuarias.

Preservar la coherencia y la trazabilidad de los resultados. Si la IA repite un análisis, es deseable que obtenga resultados coherentes. El humano debe vigilar la reproducibilidad: ejecutar dos veces el mismo pipeline sobre los mismos datos debería conducir a resultados similares; de no ser así, puede haber estocasticidad no controlada (semillas aleatorias sin fijar, cambios en particiones de entrenamiento y prueba, etc.) que conviene corregir para poder confiar en las conclusiones. El experto puede instruir a la IA

para que fije semillas aleatorias o utilice conjuntos de prueba estables, asegurando así la auditabilidad de los resultados.

En conjunto, estas prácticas se alinean con un enfoque permanente de human-in-the-loop: la IA propone soluciones y la persona experta las aprueba, refina o descarta. Este ciclo iterativo combina la velocidad de la IA con el criterio humano y tiende a producir resultados más sólidos que cualquiera de los dos por separado.

Utilizar la IA generativa como apoyo en documentación y comunicación, con revisión posterior. Otra área en la que la IA resulta especialmente útil es la generación de reportes y visualizaciones de manera automatizada (por ejemplo, código para gráficas o tablas de resultados). En los experimentos realizados, la IA produjo salidas estructuradas (como documentos en R Markdown) que facilitan la comunicación de hallazgos. Es recomendable aprovechar esta capacidad: tras el análisis, puede solicitarse a la IA que genere un informe con los resultados principales, incluya tablas comparativas de métricas por modelo o proponga figuras para presentaciones. Esto agiliza la elaboración de documentos técnicos, presentaciones o incluso borradores de artículos científicos. No obstante, es fundamental que una persona revise cuidadosamente este contenido antes de su uso formal. La IA puede expresar una idea de forma imprecisa o sin el contexto suficiente (por ejemplo, afirmar que "el modelo es 100 % preciso" sin advertir la posibilidad de sobreajuste). El experto debe editar el texto, matizar afirmaciones y asegurar que las conclusiones estén debidamente fundamentadas. En el contexto de una tesis, la IA puede apoyar en el formateo de tablas (como las tablas de resultados de modelos) o en la redacción inicial de descripciones, pero la responsabilidad de validar y adaptar el contenido al estilo académico recae en el autor o autora.

Asegurar la trazabilidad y evitar la "caja negra". Uno de los riesgos de incorporar IA generativa es aceptar sus resultados sin una comprensión adecuada del proceso que los generó, con el consiguiente riesgo de trabajar con modelos de caja negra cuyo origen y supuestos no están claros. Para mitigar este problema, es recomendable que toda interacción relevante con la IA quede registrada (por ejemplo, conservar los prompts y las respuestas, así como los cuadernos de trabajo o notebooks generados). En este estudio, los procesos de la IA se documentaron de ese modo, lo que permitió luego auditar las

decisiones adoptadas. Esta práctica debería generalizarse, tratando a la IA como un colaborador cuyo trabajo requiere versionamiento y trazabilidad, del mismo modo que el de cualquier otro miembro del equipo.

Siempre que sea posible, conviene validar de forma independiente los resultados críticos producidos por la IA. Si la IA indica que un Random Forest alcanzó 97.0 % de exactitud balanceada, es deseable que el humano realice una pequeña validación cruzada adicional o calcule manualmente la métrica a partir de las predicciones, con el fin de confirmar el valor reportado. Este tipo de contraste permite detectar errores en el código generado o malentendidos en la interpretación de métricas. Durante los experimentos, se confió en buena medida en las salidas de la IA, pero en algunas ocasiones el experto tuvo que aclarar diferencias de protocolo (validación cruzada frente a esquema hold-out) para asegurar comparaciones justas. Este doble chequeo es una práctica saludable.

También es aconsejable incorporar herramientas de interpretabilidad al pipeline de IA. Además de entrenar modelos, puede pedirse a la IA que produzca análisis de importancia de variables, gráficas basadas en SHAP u otros métodos similares, a fin de comprender el comportamiento del modelo. Esto no solo refuerza la confianza en las predicciones, sino que ayuda a detectar problemas potenciales (por ejemplo, si una variable inesperada aparece como la más importante, podría estar ocurriendo fuga de datos). La IA puede automatizar la generación de estas visualizaciones, pero la interpretación final corresponde a la persona experta.

En entornos regulados, como salud o finanzas, la trazabilidad incluye también la gestión adecuada de datos sensibles. La IA generativa debería operar, idealmente, en entornos cerrados o con datos suficientemente anonimizados, de acuerdo con las políticas de privacidad vigentes. Configurar correctamente estas restricciones y verificar su cumplimiento es responsabilidad de los equipos humanos.

Identificar escenarios donde conviene abstenerse o actuar con especial cautela. Existen situaciones en las que la IA generativa no es la herramienta más apropiada. Si el problema de datos es completamente nuevo y poco estructurado, y la exploración requiere un grado elevado de creatividad o conocimiento especializado (por ejemplo, el análisis exploratorio de variables genéticas muy específicas), la IA tenderá a ofrecer estrategias

genéricas que pueden pasar por alto matices importantes. Asimismo, cuando la calidad de los datos es muy deficiente (ruido extremo, valores faltantes no aleatorios, errores sistemáticos), suele ser preferible que una persona experta se encargue de detectar y resolver los problemas de calidad, en lugar de delegar esta tarea en una IA que aplica reglas estándar. En los casos analizados, los conjuntos de datos estaban relativamente limpios (salvo algunos faltantes), de modo que la IA no tuvo que afrontar situaciones de este tipo. En contextos más complejos es aconsejable que el humano prepare o, al menos, inspeccione de manera minuciosa los datos antes de transferirlos a la IA.

También es necesario extremar la cautela cuando se requiere una justificación formal del modelo, como ocurre en ámbitos legales o médicos. En estas circunstancias, la IA puede ser útil para identificar modelos candidatos, pero el despliegue final debería privilegiar versiones interpretables y adecuadamente documentadas. La IA generativa resulta especialmente valiosa como asistente que acelera el trabajo, más que como solución única en contextos donde se reclaman garantías de fiabilidad y explicabilidad.

A partir de estas recomendaciones puede afirmarse que la IA generativa tiene un potencial considerable para aumentar la productividad y el alcance del trabajo de las personas dedicadas a la ciencia de datos, siempre que se inserte en un marco de supervisión y colaboración humana. Diversos trabajos recientes coinciden en que la adopción de estas herramientas permite a los equipos humanos concentrarse en tareas de mayor nivel cognitivo y en actividades de mayor valor añadido, mientras la IA se ocupa de buena parte del trabajo repetitivo. Esto se traduce en equipos más eficientes y en analistas que pueden dedicar más tiempo al diseño experimental, la interpretación de resultados y la toma de decisiones estratégicas, en lugar de invertirlo en codificación manual rutinaria. La condición para que este potencial se materialice es diseñar la interacción de forma que la IA potencie al humano y no lo sustituya sin control: mantener a las personas informadas de cada paso, listas para intervenir y tratar a la IA como una herramienta sofisticada, pero no infalible. Cuando se alcanza este equilibrio, se combinan la precisión y la fiabilidad asociadas al conocimiento experto con la velocidad y amplitud de análisis que ofrece la inteligencia artificial generativa. El resultado son soluciones más rápidas, robustas y alineadas con las necesidades reales, sin perder de vista que la responsa-

bilidad última y la comprensión profunda del problema siguen recayendo en el analista humano.

Capítulo 5

Conclusiones, contribución y trabajos futuros

5.1. Conclusiones

Este trabajo comparó, de manera controlada y con trazabilidad completa, dos modos de ejecutar el ciclo CRISP-DM en problemas biomédicos de clasificación:

a. *Un flujo guiado por un analista humano, y*
b. *Un flujo asistido por un modelo de IA generativa encargado de automatizar gran parte del andamiaje técnico (carga, limpieza inicial, generación de código, búsqueda de hiperparámetros y evaluación).*

A partir del análisis cuantitativo del Capítulo 4 y la revisión cualitativa de los pasos, decisiones y resultados, pueden sintetizarse las siguientes conclusiones principales.

Primero, la hipótesis central, una ventaja del experto humano de al menos 3 % en métricas críticas, se sostiene en aquellos escenarios donde ambos enfoques siguen protocolos estrictamente emparejados. En tareas que exigen criterio experto para el preprocesamiento fino, resolución de incidencias (tipos de variables, columnas problemáticas, dependencias) y ajuste cuidadoso de ensambles, el humano transforma su juicio metodológico en una ganancia estable y medible. Esto concuerda con décadas de literatura

donde el experto destaca en contextos con ruido, desbalance extremo y estructura latente compleja.

Segundo, en datasets con alta separabilidad de clases o con señales predictivas muy fuertes (como ASD), ambos enfoques alcanzan desempeños equivalentes y cercanos al óptimo. La diferencia entre IA y humano depende más de pequeños detalles del protocolo (validación cruzada, estratificación, particiones reproducibles) que de la naturaleza humana o automática del flujo. En estos casos, la unidad metodológica del protocolo es indispensable para una lectura correcta de resultados.

Tercero, la IA generativa produce resultados competitivos con rapidez y repetibilidad, aunque con mayor variabilidad entre algoritmos. En particular, SVM, modelos de boosting y Random Forest concentraron los mejores intercambios sesgo–varianza bajo la ejecución automática, mientras que la regresión logística mostró degradación notable cuando no se aplicaron transformaciones o normalización adecuadas. La IA actúa así como un copiloto de productividad: acelera la generación de pipelines útiles, homogeniza convenciones y mejora la trazabilidad inicial; sin embargo, no sustituye la curaduría experta cuando el objetivo es maximizar confiabilidad y robustez en aplicaciones clínicas.

Cuarto, la principal ventaja del humano se manifiesta en la gobernanza del experimento: definición estricta del protocolo de evaluación (validación cruzada estratificada, semillas controladas), gestión explícita del desbalance (umbral, ponderación, remuestreo), prevención del data leakage, y reporte disciplinado de métricas relevantes (exactitud balanceada, sensibilidad, especificidad, intervalos de confianza). La IA, en contraste, sobresale en la ejecución veloz y sin fatiga de instrucciones bien definidas.

Quinto, la perspectiva operativa resultante es de complementariedad. El mejor desempeño surge al combinar:

- *La velocidad y reproducibilidad de la IA generativa para construir bases funcionales, y*
- *El criterio experto para cerrar brechas metodológicas, reforzar la validez del pipeline y optimizar el rendimiento final.*

Este enfoque híbrido reduce ciclos de iteración, disminuye errores de configuración y

aumenta la confiabilidad del entregable.

5.2. Implementación de agentes de IA en el ciclo CRISP–DM

Los resultados indican que la IA generativa puede desempeñar un papel todavía más eficaz cuando se encapsula en agentes especializados capaces de interactuar directamente con datos, artefactos intermedios, métricas y modelos. Esta arquitectura extiende la IA más allá de un asistente que sólo genera código, hacia un sistema autónomo que ejecuta acciones, verifica resultados y adapta su conducta según los objetivos del proyecto.

A continuación se describen los componentes fundamentales para integrar agentes de IA en CRISP–DM:

5.2.1. Agentes con acceso estructurado a datos y metadatos

Un agente puede operar de forma segura mediante:

- *Acceso controlado a tablas (lectura/escritura delimitada),*
- *Esquemas declarativos de datos (tipos, rangos, codificaciones),*
- *Consultas parametrizadas,*
- *Catálogos de metadatos (diccionario de variables, unidades clínicas, dominios válidos).*

*Esto permite que la IA no sólo genere código, sino que **verifique** el estado real de los datos, detecte erratas, inconsistencias, valores fuera de rango o violaciones de esquema antes de iniciar el modelado.*

5.2.2. Agentes capaces de ejecutar código, medir resultados y modificar su pipeline

Un agente puede controlar un entorno de ejecución (R, Python, SQL), correr experimentos, registrar métricas y decidir los siguientes pasos. Este bucle de acción–observación–decisión transforma el trabajo analítico en:

- *Ciclos automáticos de prueba de modelos,*
- *Ajuste iterativo de hiperparámetros,*
- *Selección de mejores configuraciones según criterios clínicos,*
- *Reentrenamiento cuando detecta deriva o degradación del modelo.*

Con esto, la IA deja de ser un simple generador de código y se convierte en un ejecutor autónomo que ve sus propios resultados y aprende a partir de ellos.

5.2.3. Agentes para evaluación y auditoría

Los agentes pueden:

- *Generar automáticamente validaciones cruzadas,*
- *Comparar métricas entre modelos,*
- *Calcular intervalos de confianza,*
- *Estructurar reportes replicables,*
- *Verificar umbrales clínicos mínimos,*
- *Detectar inestabilidad o sobreajuste.*

Esta actuación asegura que la IA mantenga disciplina metodológica sin intervención humana constante.

5.2.4. Agentes orquestadores de pipelines completos

Un nivel más avanzado consiste en un agente maestro que coordina:

- *un agente de preprocesamiento,*
- *un agente de modelado,*
- *un agente de evaluación,*
- *un agente de documentación,*
- *un agente de despliegue y monitoreo.*

Este sistema convierte CRISP–DM en un flujo altamente automatizado, donde cada módulo IA se especializa y reporta sus decisiones para trazabilidad.

5.2.5. Comunicación Humano–IA basada en criterios de decisión

Los agentes pueden:

- *explicar por qué escogieron un modelo,*
- *justificar umbrales,*
- *advertir sobre riesgos de especificidad baja o sensibilidad insuficiente,*
- *sugerir mejoras en ingeniería de variables.*

Esto permite una interacción más cercana al trabajo entre analista y asistente técnico, donde la IA no sólo ejecuta sino que también argumenta y reporta.

5.2.6. Agentes para monitoreo post-despliegue

- *En producción, los agentes permiten:*
- *seguimiento de deriva de datos,*
- *cálculo periódico de sensibilidad/especificidad en nuevos lotes,*

- *detección temprana de condiciones anómalas,*
- *recomendaciones de reentrenamiento,*
- *activación de alertas cuando la distribución de entrada se desvía del entrenamiento.*

Así, la IA participa activamente en el ciclo de vida del modelo, no sólo en su desarrollo inicial.

El uso de agentes IA convierte a la IA generativa en un elemento operativo del ciclo CRISP–DM: no sólo produce código, sino que observa, evalúa, dialoga, corrige y ejecuta. Este enfoque impulsa una transición desde la automatización parcial hacia un ecosistema híbrido en el que la IA es coautora del pipeline analítico y el humano supervisor estratégico.

5.3. Limitaciones del estudio

A pesar de los hallazgos robustos obtenidos en la comparativa entre el experto humano y la inteligencia artificial generativa, este trabajo presenta limitaciones inherentes a su diseño experimental y al estado actual de la tecnología, las cuales deben considerarse al extrapolar las conclusiones.

Dependencia del modelo y volatilidad tecnológica El estudio se centró exclusivamente en el modelo GPT-4o de OpenAI. Dada la velocidad de evolución en el campo de los Grandes Modelos de Lenguaje (LLMs), los resultados presentados constituyen una 'instantánea' de las capacidades disponibles en el momento de la experimentación. Es posible que modelos posteriores (como GPT-5 o versiones avanzadas de Claude y Gemini) superen las deficiencias detectadas aquí, particularmente en el manejo de sesgos y razonamiento clínico, o que modelos más pequeños presenten brechas de rendimiento diferentes.

Sesgo del experto único. *La referencia de desempeño humano ('Gold Standard') fue establecida por un único experto en ciencia de datos utilizando lenguaje R. Si bien se siguieron protocolos estrictos y metodologías estándar, las decisiones de modelado (como*

la elección específica de hiperparámetros o la estrategia de ingeniería de características) contienen inevitablemente un componente subjetivo basado en la experiencia individual. Un experto con diferente formación o un equipo de analistas podría haber tomado decisiones distintas que alterarían el margen de diferencia respecto a la IA.

Restricción al dominio de datos tabulares estructurados. *La investigación se limitó a tres conjuntos de datos del ámbito de la salud (Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set), todos ellos de naturaleza tabular y estructurada. Aunque se varió la dimensionalidad y el balance de clases, el estudio no evaluó el desempeño de la IA en datos no estructurados (imágenes médicas, notas clínicas en texto libre) ni en otros dominios fuera de la biomedicina (como finanzas o marketing), donde la IA podría exhibir comportamientos y alucinaciones diferentes.*

Naturaleza estática de la interacción (*Prompt Engineering*) *La metodología se basó en prompts estandarizados y diseñados cuidadosamente para cada etapa del ciclo CRISP-DM. El rendimiento de la IA generativa es altamente sensible a la formulación del prompt; variaciones en la instrucción (por ejemplo, few-shot prompting vs. zero-shot) podrían haber resultado en un código de mayor o menor calidad. Por tanto, los resultados reflejan la capacidad de la IA bajo una estrategia de prompts específica, no necesariamente su límite teórico máximo.*

Ausencia de ejecución autónoma en tiempo real En el diseño experimental, la IA generó el código, pero la ejecución y la validación final recayeron en el entorno controlado por el humano. La IA no tuvo acceso a un entorno de ejecución en tiempo real para corregir sus propios errores (tales como fallos en dependencias de librerías o dimensiones de matrices) durante la generación. Esto limita la evaluación de la IA como un agente totalmente autónomo, evaluándola más bien como un asistente de generación de código que requiere un operador humano para el despliegue.

5.4. Contribución de la investigación

Las contribuciones de esta tesis se organizan en cuatro planos, conceptual, metodológico, empírico y práctico,con la finalidad de proporcionar un marco reutilizable para equipos que deseen integrar IA generativa y agentes automatizados en el ciclo CRISP-DM.

Contribuciones conceptuales

Marco comparativo humano–IA dentro de CRISP-DM. *Se formaliza una lectura paralela de las fases del proceso, identificando en qué etapas la IA generativa puede automatizar tareas y en cuáles la intervención humana sigue siendo indispensable.*

Taxonomía de riesgos en flujos automatizados. *Se documentan fallas típicas en la ejecución automática (tipificación incorrecta de columnas, supuestos implícitos de modelos, manejo inadecuado de clases raras, dependencias de librerías) y se proponen checks tempranos para su detección.*

Justificación de métricas adecuadas en los problemas analizados. *A partir de la naturaleza de los tres datasets utilizados, se explicita por qué métricas como exactitud balanceada, sensibilidad y especificidad resultan más informativas que la precisión global, integrando además la calibración de probabilidades como parte del proceso evaluativo.*

Contribuciones metodológicas

Protocolo de evaluación unificado y reproducible. *Se establece un esquema basado en validación cruzada estratificada con semillas controladas, reporte de varianza entre pliegues y trazabilidad estricta de cada decisión analítica.*

Generación automatizada de artefactos ejecutables mediante IA generativa. *Se desarrolló un mecanismo para producir documentos R Markdown completos y ejecutables,*

integrando código, análisis, visualizaciones y narrativa técnica. Este proceso se apoyó en Jupyter Notebooks, agentes personalizados y flujos controlados que transforman insumos del usuario en documentos reproducibles.

Diseño de agentes autónomos de apoyo analítico. *Se creó un agente conversacional accesible por WhatsApp capaz de realizar una exploración rápida del conjunto de datos, solicitar los parámetros esenciales y devolver un archivo R Markdown listo para ejecutarse. Este agente funciona como interfaz de campo para facilitar el uso aplicado de CRISP-DM.*

Prompts estandarizados para cada fase de CRISP-DM. *Se desarrollaron instrucciones formales para guiar a la IA generativa en comprensión de datos, preparación, modelado, evaluación y documentación, reduciendo la variabilidad asociada a instrucciones informales.*

Lista de robustez para procesos híbridos humano–IA. *Se presenta un conjunto de verificaciones para prevenir data leakage, asegurar tipologías correctas, registrar versiones, documentar semillas y mantener reproducibilidad de extremo a extremo.*

Contribuciones empíricas

Comparativa sistemática entre IA generativa y experto humano en tres problemas biomédicos concretos. *Se identifican patrones consistentes en los datasets analizados: los modelos SVM, boosting y bosques aleatorios alcanzan los mejores resultados; la regresión logística requiere preprocesamiento cuidadoso; y los ensambles construidos por el experto ofrecen ventajas en escenarios de dificultad intermedia.*

Análisis por régimen de dificultad *Se muestra que ambos enfoques alcanzan desempeños altos en datasets altamente separables; que la IA generativa puede superar al humano en problemas con alta dimensionalidad y abundancia de datos; y que el humano obtiene ventajas claras cuando el ajuste fino es determinante para el rendimiento final.*

Contribuciones prácticas

- *Integración de un flujo automatizado para generación de documentos analíticos. Se desarrolló un mecanismo funcional mediante agentes conversacionales (incluyendo un agente accesible por WhatsApp) capaz de solicitar parámetros esenciales al usuario y producir un archivo R Markdown ejecutable, estructurado según las etapas de CRISP-DM.*
- *Estructuración de un proceso reproducible desde la exploración hasta la generación de artefactos ejecutables. Se definieron procedimientos para transformar la entrada del usuario en documentos analíticos consistentes que combinan código, visualizaciones y narrativa técnica.*
- *Plantillas estandarizadas de R Markdown generadas por IA. Se diseñó una estructura fija que la IA puede completar de manera fiable, asegurando uniformidad entre proyectos y facilitando tanto la revisión manual como la extensión del flujo hacia nuevas tareas analíticas.*

5.5. Trabajos futuros

A partir de los hallazgos y las limitaciones identificadas en esta investigación, se proponen las siguientes líneas de trabajo para profundizar en la integración de la inteligencia artificial generativa en la ciencia de datos:

Evaluación comparativa con múltiples arquitecturas de LLMs. *Dado que este estudio se centró exclusivamente en GPT-4o, un paso natural es replicar la metodología experimental utilizando otros modelos de vanguardia, como Claude 3.5 Sonnet, Gemini 1.5 Pro o modelos de código abierto como Llama 3. Esto permitiría determinar si los sesgos de çaja negraz la falta de sensibilidad clínica observados son inherentes a la tecnología actual de LLMs o si varían significativamente entre distintas arquitecturas y entrenamientos.*

Extensión a datos no estructurados y multimodales *El presente trabajo se limitó a datos tabulares estructurados. Futuras investigaciones deberían explorar el desempeño de la IA en el ciclo CRISP-DM aplicado a datos no estructurados, tales como imágenes médicas (radiografías, resonancias), señales biomédicas crudas (series de tiempo de EEG sin preprocesar) o notas clínicas en texto libre. Evaluar la capacidad de modelos multimodales para integrar estas fuentes de información heterogéneas representaría un avance significativo hacia una IA más holística en salud.*

Implementación de técnicas de RAG (Retrieval-Augmented Generation). *Para mitigar la alucinación y mejorar la toma de decisiones contextuales, se propone investigar el impacto de integrar sistemas de Generación Aumentada por Recuperación (RAG). Conectar al agente de IA a bases de conocimiento médico actualizadas, guías de práctica clínica y documentación técnica de librerías de R podría mejorar la justificación de los modelos y reducir la brecha de conocimiento experto detectada en el caso de diabetes.*

Desarrollo y prueba de campo de los agentes propuestos *Esta tesis delineó teóricamente la arquitectura de agentes autónomos para el ciclo CRISP-DM. Un trabajo futuro inmediato es la implementación técnica de estos agentes y su despliegue en un entorno controlado de producción (piloto). Esto permitiría medir métricas operativas reales, como la reducción en horas-hombre, la latencia en la entrega de modelos y la usabilidad de la interacción humano-IA por parte de equipos de datos reales.*

Investigación sobre explicabilidad automática (Auto-XAI). *Finalmente, se sugiere desarrollar módulos específicos donde la IA no solo genere el modelo predictivo, sino que autogenere explicaciones de sus decisiones (utilizando valores SHAP, LIME o contrafactuales) y las presente en lenguaje natural al experto humano. Investigar si estas explicaciones automáticas mejoran la capacidad del humano para detectar sesgos algorítmicos constituiría un aporte valioso a la gobernanza de la IA.*

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	*Exploración de datasets usando IA generativa frente a un experto humano en lenguaje R*
Autor:	*Jorge Fred Alvarez Salaya*
ORCID:	*https:// orcid.org/ 0009-0008-7309-098X*
Resumen:	*Este trabajo de investigación analiza el papel de la inteligencia artificial generativa en la ciencia de datos, comparando el desempeño de GPT-4o con el de un experto humano a lo largo del ciclo de vida CRISP-DM. El trabajo se inscribe en un contexto donde el volumen de datos y la presión por automatizar tareas analíticas hacen cada vez más relevante la posibilidad de delegar parte del proceso a modelos de lenguaje de gran tamaño, sin perder rigor ni calidad en problemas reales del ámbito de la salud.* *La metodología se basa en tres conjuntos de datos médicos (Autistic Spectrum Disorder Screening Data for Children, Epileptic Seizure Recognition Data Set y Diabetes Data Set), en los que se ejecuta el mismo flujo de trabajo por duplicado: por un lado, mediante scripts en R desarrollados por un experto, y por otro, a partir de indicaciones dirigidas a GPT-4o, que generan documentos R Markdown reproducibles. En ambos casos se entrenan los mismos algoritmos de aprendizaje automático y se comparan sus resultados con métricas estándar.* *Los resultados muestran que GPT-4o puede igualar o incluso superar al humano en problemas con estructura más clara (autismo y epilepsia), pero ofrece un rendimiento sensiblemente inferior y más inestable en el caso de diabetes. A partir de ello, la tesis concluye que la IA generativa no sustituye al experto, aunque sí puede funcionar como asistente que automatiza tareas repetitivas y acelera la experimentación, dentro de esquemas híbridos donde el control y la decisión final siguen recayendo en la persona especialista.*
Palabras clave:	*IA generativa; ciencia de datos; CRISP-DM; GPT-4o; agentes inteligentes.*
Referencias citadas:	*En la siguiente página se muestran las referencias.*

Bibliografía

Abbasov, A. (2023). Forecasting Hourly Electricity Prices in European Nations: Utilizing Machine Learning and Deep Learning Techniques for Sustainable Energy Decision-Making. *Journal of Harbin Engineering University*, *44*(7), 2383-2392.

Alaminos-Fernández, A. F. (2023). *Árboles de decisión en R con Random Forest* [Ciencia Social Computacional]. Limencop S.L. https://rua.ua.es/dspace/bitstream/10045/133067/1/Random_Forest_en_la_Investigacion_Social.pdf

Al-Haija, Q. A. (2022). Exploration of Tools for Data Science. En *Data Science with Semantic Technologies* (pp. 31-69). John Wiley Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119865339.ch2

Aljanabi, M., Ghazi, M., Ali, A. H., Abed, S. A., et al. (2023). ChatGPT: open possibilities. *Iraqi Journal For Computer Science and Mathematics*, *4*(1), 62-64.

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, *27*(1), 17-21. https://doi.org/10.1080/00031305.1973.10478966

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT)*, 144-152. https://doi.org/10.1145/130385.130401

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215-232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303-314. https://doi.org/10.1007/BF02551274

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, *55*(10), 78-87. https://doi.org/10.1145/2347736.2347755

El-Din, D. M., Hassanien, A. E., & Hassanien, E. E. (2020). Challenges of Big Data Visualization in Internet-of-Things Environments. En *Advances in Intelligent Systems and Computing* (pp. 873-885). Springer Singapore. https://doi.org/10.1007/978-981-15-1286-5_76

Emmert-Streib, F., Moutari, S., & Dehmer, M. (2016). The Process of Analyzing Data is the Emergent Feature of Data Science. *Frontiers in Genetics*, *7*. https://doi.org/10.3389/fgene.2016.00012

Fang, L., Lee, G.-G., & Zhai, X. (2023). Using GPT-4 to Augment Unbalanced Data for Automatic Scoring.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI [Last revised: August 7, 2023]. *Business & Information Systems Engineering*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4443189

Fix, E., & Hodges, J. L. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties* (inf. téc. N.º Project 21-49-004, Report No. 4). USAF School of Aviation Medicine. Randolph Field, Texas.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

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

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2672-2680.

Govindarajan, M. (2020). Challenges in Big Data Analysis. En *Challenges in Big Data Analysis* (pp. 577-585). IGI Global. https://doi.org/10.4018/978-1-7998-3479-3.ch041

Gracla, S., Bockelmann, C., & Dekorsy, A. (2022). On the Importance of Exploration for Real Life Learned Algorithms, 1-5. https://doi.org/10.1109/SPAWC51304.2022.9834009

Hassan, M. M., Knipper, A., & Santu, S. K. K. (2023). ChatGPT as your Personal Data Scientist. https://doi.org/10.48550/arxiv.2305.13657

Hosmer, D. W., & Lemeshow, S. (1980). Goodness of Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics - Theory and Methods*, *A9*, 1043-1069. https://doi.org/10.1080/03610928008827941

Human Uncertainty and Ranking Error – The Secret of Successful Evaluation in Predictive Data Mining. (2017). *arXiv: Human-Computer Interaction*.

Hüsing, S. (2021). Epistemic Programming - An Insight-Driven Programming Concept for Data Science. https://doi.org/10.1145/3488042.3490510

IBM. (2025). *¿Qué es un árbol de decisión?* [Consultado el 13 de abril de 2025]. IBM. https://www.ibm.com/es-es/think/topics/decision-trees

Javier. (2025). *Modelo de regresión logística ¿Qué es y para qué sirve?* [Consultado el 13 de abril de 2025]. IAY Programación. https://www.iayprogramacion.com/modelo-de-regresion-logistica/

Jeha, G. M., Qiblawi, S., Jairath, N., Sable, K., LeBlanc, K., Aylward, J., & Xu, Y. G. (2023). ChatGPT and Generative Artificial Intelligence in Mohs Surgery: A New Frontier of Innovation. *Journal of Investigative Dermatology*, *143*(11), 2105-2107. https://doi.org/https://doi.org/10.1016/j.jid.2023.05.018

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Karthika, N., Sheela, J., & Janet, B. (2022). A Brief Introduction and Importance of Data Science. En *Data Science with Semantic Technologies* (pp. 1-30). John Wiley Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119865339.ch1

Keim, D. A., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. En A. Kerren, J. T. Stasko, J.-D. Fekete & C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives* (pp. 154-175, Vol. 4950). Springer. https://doi.org/10.1007/978-3-540-70956-5_7

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lee, R. (2020). *Artificial Intelligence in Daily Life*. Springer Nature Singapore. https://books.google.com.mx/books?id=NOH4DwAAQBAJ

Ma, P., Ding, R., Wang, S., Han, S., & Zhang, D. (2023). Demonstration of InsightPilot: An LLM-Empowered Automated Data Exploration System.

Malik, A., Thevisuthan, P., & De Silva, T. (2022). Artificial Intelligence, Employee Engagement, Experience, and HRM. En A. Malik (Ed.), *Strategic Human Resource Management and Employment Relations: An International Perspective* (pp. 171-184). Springer. https://doi.org/10.1007/978-3-030-90955-0_16

Milligan, I. (2022). *The Transformation of Historical Research in the Digital Age*. Cambridge University Press. https://doi.org/10.1017/9781009026055

Morris, M. R. (2023). Scientists' Perspectives on the Potential for Generative AI in their Fields.

OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Palazhchenko, Y., Shendryk, V., & Shendryk, S. (2023). *Digital Twins Data Visualization Methods. Problems of Human Interaction: A Review* (I. Karabegovic, A. Kovačević & S. Mandzuka, Eds.). Springer Nature Switzerland.

Prentice, R. L., & Pyke, R. (1979). Logistic Disease Incidence Models and Case-Control Studies. *Biometrika*, *66*(3), 403-411. https://doi.org/10.1093/biomet/66.3.403

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/BF00116251

Rabanal, I. M. (2024). Interpretabilidad Global de Modelos de Machine Learning para la Gestión de Riesgos Financieros: Un Enfoque Agnóstico al Modelo [Trabajo de Fin de Grado, Facultad de Ciencias Económicas y Empresariales]. https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/79200/TFG_Moreno_Rabanal_Ines.pdf

Rodríguez Yunta, L. (2017). Difusión y evaluación de la investigación histórica en la era digital: revistas españolas y bases de datos = Dissemination and evaluation of historical reseach in the Digital Age: Spanish journals and databases. *Estudios Humanísticos. Historia*, (15), 205-238. https://doi.org/10.18002/ehh.v0i15.5048

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, *27*(75). https://doi.org/10.1186/s13054-023-04380-2

Samuelson, P. (2023). Generative AI meets copyright. *Science*, *381*(6654), 158-161. https://doi.org/10.1126/science.adi0656

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, *2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Stadlmann, C., & Zehetner, A. (2021). Human Intelligence Versus Artificial Intelligence: A Comparison of Traditional and AI-Based Methods for Prospect Generation. En Á. Rocha, J. L. Reis, M. K. Peter, R. Cayolla, S. Loureiro & Z. Bogdanović (Eds.), *Marketing and Smart Technologies* (pp. 11-22). Springer Singapore.

Syafganti, I. (2018). Digital Transformation, Big Data and Research Landscape in Digital Communication. *Jurnal Komunikasi Ikatan Sarjana Komunikasi Indonesia*, *3*(2). https://doi.org/10.25008/JKISKI.V3I2.220

Tang, N., Yang, C., Fan, J., Cao, L., Luo, Y., & Halevy, A. (2023). VerifAI: Verified Generative AI.

Thabtah, F. (2017). Autistic Spectrum Disorder Screening Data for Children [[Dataset]]. https://doi.org/10.24432/C5659W

Urs, S. R., & Minhaj, M. (2023). Evolution of data science and its education in iSchools: An impressionistic study using curriculum analysis. *Journal of the Association for Information Science and Technology*, *74*(6), 606-622. https://doi.org/https://doi.org/10.1002/asi.24649

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), Article 25. https://doi.org/10.2202/1544-6115.1309

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Wang, Z., Li, Y., Wu, J., Soon, J., & Zhang, X. (2023). FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis.

Wilcox, R. R. (2009). *Basic Statistics: Understanding Conventional Methods and Modern Insights*. Oxford University Press. https://www.ctanujit.org/uploads/2/5/3/9/25393293/_basic_statistical_theory.pdf

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, *5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

World Bank. (2023). Generative Artificial Intelligence [License: CC BY-NC 3.0 IGO]. http://hdl.handle.net/10986/39959

Zhang, K., Liu, S., & Xiong, M. (2022). Changes from Classical Statistics to Modern Statistics and Data Science.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*.

Zohny, H., McMillan, J., & King, M. (2023). Ethics of generative AI. *Journal of Medical Ethics*, *49*(2), 79-80. https://doi.org/10.1136/jme-2023-108909