INTRODUCCIÓN

Este libro nació de una motivación personal y académica: la convicción de que un ingeniero que entienda, transforme y otorgue valor agregado a los datos está mejor preparado que aquel que se limita al entrenamiento de modelos. Desde que inicié mis primeros trabajos en *machine learning* y ciencia de datos, he podido presenciar la rápida evolución del campo y, al mismo tiempo, los vacíos conceptuales que aún persisten en la formación de muchos profesionales. Con frecuencia, los estudiantes se maravillan con los resultados de los modelos de aprendizaje automático, pero desconocen las estructuras invisibles que los hacen posibles: la ingeniería de datos detrás de cada tipo de información, su calidad, su trazabilidad y su significado.

Fue precisamente desde esa observación, y desde más de una década de trabajo académico e investigativo en procesamiento de señales, visión por computador y detección de audio sintético, que surgió la idea de este libro. Su propósito es guiar a los futuros ingenieros hacia una comprensión integral del ciclo de vida de la ciencia de datos, combinando el rigor técnico con la curiosidad científica. No se trata solo de aprender librerías o ejecutar modelos, sino de pensar con datos, de cuestionarlos, de entender su contexto y su propósito.

El enfoque de este libro se enmarca en la metodología de Aprendizaje Basado en Proyectos (ABP), lo que significa que cada capítulo propone retos y proyectos que conectan la teoría con la práctica investigativa. De esta manera, el aula se convierte en un laboratorio de exploración, donde el estudiante no solo reproduce experimentos, sino que diseña los suyos, analiza resultados, documenta hallazgos y aprende a comunicar evidencia científica.

A lo largo de los capítulos, el lector recorrerá un camino progresivo que inicia con los fundamentos conceptuales de la ciencia e ingeniería de datos y culmina con un caso aplicado de alto impacto: la detección de audio sintético mediante modelos de aprendizaje profundo. El primer y segundo capítulo introduce los pilares teóricos de la ciencia de datos, la diferencia entre los

enfoques *model-centric* y *data-centric*, y la importancia de la reproducibilidad y la evaluación ética. El tercer capítulo profundiza en la ingeniería de datos estructurados y la construcción de *pipelines* reproducibles, con ejemplos reales de optimización y compresión de modelos. El cuarto explora las series de tiempo y el procesamiento de señales, mostrando cómo los datos pueden narrar comportamientos, por ejemplo, en el sector bancario. Finalmente, el quinto capítulo integra todas las dimensiones anteriores en un proyecto de detección de voz sintética, donde confluyen la ingeniería de datos, el procesamiento espectral y la ética en la era de los *deepfakes*.

Los temas y experimentos de este libro se nutren de investigaciones publicadas entre 2019 y 2025, fruto de colaboraciones científicas en torno a la eficiencia de modelos de aprendizaje profundo, la interpretabilidad de las redes neuronales y la detección de falsificaciones de voz e imagen. Trabajos como *Deep4SNet* y *FlexiPrune*, sirven de base experimental y metodológica para los capítulos, aportando ejemplos reales de cómo la ingeniería de datos se convierte en una herramienta de investigación aplicada. Esta integración permite al lector no solo estudiar los conceptos, sino entender cómo se construyen y validan en la práctica científica.

En el contexto profesional, la ingeniería de datos se ha consolidado como una de las competencias más valoradas en sectores como las telecomunicaciones, multimedia, la banca, la seguridad digital y el Internet de las Cosas. Dominar el ciclo de vida del dato, desde su captura hasta su modelado, permite construir sistemas de inteligencia artificial eficientes, confiables y éticamente responsables. Por ello, este libro también es una invitación a pensar la ingeniería de datos no solo como una disciplina técnica, sino como una práctica con responsabilidad social.

El ecosistema tecnológico que lo acompaña incluye las principales librerías de Python empleadas en ciencia e ingeniería de datos, como *pandas*, *numpy*, *scikit-learn*, *tensorflow*, *torch*, *librosa*, *torchaudio*, entre otras. Cada una se presenta no como una herramienta aislada, sino como parte de una arquitectura integral de trabajo que el lector aprenderá a ensamblar, documentar y optimizar. Los ejemplos se basan en conjuntos de datos reales provenientes tanto de entornos industriales como de investigación académica, lo que facilita la conexión entre teoría y práctica.

Al finalizar este recorrido, el lector no solo habrá adquirido habilidades técnicas, sino una forma distinta de mirar los datos: con rigor, con curiosidad y con conciencia. Comprenderá que cada modelo de aprendizaje automático es tan sólido como el conjunto de decisiones que lo preceden, y que, en esa cadena, desde la adquisición hasta la inferencia, reside el verdadero valor de la ingeniería. En última instancia, este libro es una invitación a explorar, a cuestionar y a crear, con mente abierta, manos en el código y espíritu de investigador.