UNED



# ESTADÍSTICA APLICADA CON INTELIGENCIA ARTIFICIAL

Alfonso García Pérez

# *Estadística aplicada con inteligencia artificial*

ALFONSO GARCÍA PÉREZ

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

ESTADÍSTICA APLICADA CON INTELIGENCIA ARTIFICIAL
0105022CT01l01 • *Ciencias*

Esta publicación ha sido evaluada por pares, método doble ciego,
por expertos ajenos a la universidad.

ISBN electrónico: 978-84-362-8325-9

Primera edición: octubre de 2026

© Alfonso García Pérez

© Universidad Nacional de Educación a Distancia

**www.uned.es/publicaciones**

Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del *copyright*, bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos.

***Aquí*** podrá encontrar información adicional y actualizada de esta publicación

Dedicado a mi madre y a mi padre,
que me lo dieron todo.
Y a mis cinco nietas,
*Almudena*, *Fátima*, *Claudia*,
*Isabel* y *Blanca*,
que son el futuro y me han dado la
fuerza necesaria para escribir este libro.

# Prólogo

Muchas técnicas estadísticas, como la regresión lineal, requieren de la ejecución de tests de hipótesis; por ejemplo, en la regresión lineal se requiere contrastar si las covariables consideradas son o no significativas, pero el estadístico de contraste suele tener un error de muestreo que depende del tamaño muestral en su denominador, lo que hace que, en cuanto el tamaño muestral sea muy grande, se suela concluir que todas las covariables consideradas son significativas.

Es decir, la estadística que tradicionalmente utilizamos suele necesitar una modificación en cuanto el número de variables o el tamaño muestral sean muy elevados.

La Inteligencia Artificial proporciona las herramientas necesarias para poder seguir utilizando la Inferencia Estadística en esas ocasiones. Para ello se divide la muestra en dos grupos: *Muestra de Entrenamiento* y *Muestra Experimental*, se aplican las técnicas estadísticas a la primera y se validan en la segunda muestra, midiendo la calidad del análisis mediante la *Matriz de Confusión*.

Además, existen técnicas propias de la Inteligencia Artificial, las cuales se han ido desarrollado por esta gran cantidad de datos manejados y la gran potencia de cálculo de los ordenadores. Nos referimos al *Machine Learning* y a las *Redes Neuronales*.

Se podría decir que este libro es una guía de cómo ejecutar Estadística Aplicada utilizando técnicas de Inteligencia Artificial.

Aunque sería posible estudiar estos métodos con otros paquetes estadísticos, nosotros hemos elegido el software gratuito R. Por ejemplo, en ocasiones se suele recomendar Python pero no hay ninguna razón ni técnica que haga preferir a este último en relación a R, por lo que mantenemos nuestra recomendación de elegir R.

Si el lector quiere replicar los comandos utilizados con R en este libro, deberá copiarlos sin utilizar el símbolo $>$ o el símbolo $+$ .

En ocasiones se citarán los textos *Estadística Básica con R* (EBR), *Estadística Aplicada: Conceptos Básicos* (CB), *Estadística Aplicada Robusta con R* (EARR), *Análisis Estadístico de Datos Espaciales con QGIS y R* (AEQR), *Estadística Aplicada Avanzada con R* (EAAR), *Problemas Básicos de Estadística Aplicada* (PBEA) y *Problemas Avanzados de Estadística Aplicada* (PAEA), todos del mismo autor de este libro y publicados en la editorial UNED. Su cita permite descargar parte del contenido de este libro e indica que, en principio, el lector debería de tener ciertos conocimientos básicos de Estadística y de R. Si no los tiene, le recomendamos leer EBR.

En este texto aparecerán referencia de datos, que se incorporan con la ruta `e:` la cual, lógicamente, deberá cambiar si usted utiliza otra ruta.

Si quiere recibir los ficheros de datos que utilizamos en el libro, así como los comandos de R, además de problemas resueltos en un Suplemento, e incluso comentarios o aclaraciones del autor, puede ponerse en contacto con él a través del correo electrónico

`estadistica.aplicada.robusta@gmail.com`

Alfonso García Pérez
Catedrático de Estadística e I.O.
UNED

# Índice

**1. Introducción a la Inteligencia Artificial: Técnicas Machine Learning 11**
1.1. Introducción . . . 11
1.2. Árboles de Clasificación y Regresión (*Classification and Regression Trees, CARTs*) . . . 13
1.2.1. Construcción de un Árbol . . . 13
1.2.2. Árboles óptimos . . . 21
1.3. Bosques Aleatorios (*Random Forests, RFs*) . . . 25
1.4. Árboles de Regresión Lógica . . . 28
1.5. Regresión a Trazos Adaptativa Multivariante (*Multivariate Adaptive Regression Splines, MARS*) . . . 30
1.6. Lecturas Recomendadas . . . 33

**2. k-vecinos más cercanos: Clasificación y Predicción 35**
2.1. Introducción . . . 35
2.2. Matriz de Confusión . . . 36
2.3. Método k vecinos más cercanos . . . 37
2.4. Lecturas Recomendadas . . . 38

**3. Regresión Lineal con técnicas Machine Learning 39**
3.1. Introducción . . . 39
3.2. Regresión Lineal con Prueba y Error . . . 40
3.3. Tamaño muestral muy elevado . . . 42
3.4. Sobre-ajuste . . . 43
3.5. Lecturas Recomendadas . . . 45

**4. Selección de Predictores: Regresión Lasso, Ridge y Elastic Net 47**
4.1. Introducción . . . 47
4.2. Regresión Ridge . . . 48
4.3. Regresión Lasso . . . 50

4.4. Elastic Net . . . . . 51
4.5. Lecturas Recomendadas . . . . . 51

**5. Regresión Logística con técnicas Machine Learning 53**
5.1. Introducción . . . . . 53
5.2. Lecturas Recomendadas . . . . . 56

**6. Deep Learning: Redes Neuronales Artificiales 57**
6.1. Introducción . . . . . 57
6.1.1. Formulación matemática de una Red Neuronal . . . . . 58
6.2. Redes Neuronales con R: librerías nnet y neuralnet . . . . . 59
6.3. Redes neuronales e inteligencia artificial . . . . . 61
6.4. Operador tubería . . . . . 63
6.5. Redes Neuronales Convolucionales . . . . . 63
6.5.1. Redes Neuronales Convolucionales con R . . . . . 66
6.6. Lecturas Recomendadas . . . . . 73

# Capítulo 1

# Introducción a la Inteligencia Artificial: Técnicas Machine Learning

## 1.1. Introducción

Comenzaremos el estudio de las técnicas de Inteligencia Artificial, a veces denominada simplemente con su acrónimo IA, estudiando en este capítulo *Machine Learning*, es decir, *Aprendizaje Automático* o *Maestro en determinar Patrones*, que es una disciplina del campo de la Inteligencia Artificial que dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo) mediante el uso de modelos a través de algoritmos.

Estos datos, utilizados con este propósito de la determinación de patrones, se denominan *Datos de Aprendizaje o Entrenamiento*. A Machine Learning se la denomina también *Statistical Learning*; esta denominación procede del campo de la Estadística mientras que *Machine Learning* procede la informática.

Es interesante destacar que el ordenador es el que crea los algoritmos basados en los datos de entrenamiento que se utilizan para calibrar el modelo de aprendizaje automático. Con base los datos de entrenamiento, el algoritmo encuentra las reglas subyacentes internamente.

Este aprendizaje permite a los ordenadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados.

Cuando los datos de aprendizaje no han sido estructurados ni etiquetados por nosotros, este método de creación del modelo se denomina *Deep Learning* (*Aprendizaje Profundo*). Este tipo de aprendizaje automático se realiza con el uso de las denominadas *Redes Neuronales Artificiales*, *Neural Networks*.

El término se utilizó por primera vez en 1959, cuando los ordenadores eran muy limitados. Dada a la gran capacidad cálculo de los ordenadores en la actualidad y a la gran cantidad de datos que puede manejar, esta disciplina ha ganado relevancia en los últimos tiempos. Las técnicas de Aprendizaje Automático son, de hecho, una parte fundamental del denominado *Big Data Analysis*.

Los algoritmos de Machine Learning, en ocasiones abreviado como ML, se dividen en dos categorías:

**Aprendizaje supervisado**: Tiene como propósito predecir una variable observada a partir de otras variables relacionadas. Estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. Un ejemplo es un detector de spam que etiqueta un e-mail como spam o no, dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.).

Los algoritmos de Aprendizaje Supervisado se dividen en Algoritmos de Regresión (cuando la variable predicha es numérica) y Algoritmos de Clasificación (cuando la variable predicha es categórica). Son Algoritmos de Regresión, la Regresión Lineal y el Random Forest.

Son Algoritmos de Clasificación la Regresión Logística y los Árboles de Decisión.

**Aprendizaje no supervisado**: Estos algoritmos no cuentan con un conocimiento previo. Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna manera. Por ejemplo, en el campo del marketing se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas.

Dada la reciente implantación de las técnicas estudiadas en este libro, en muchas ocasiones hemos mantenido la denominación anglosajona para su fácil identificación. También queremos decir que los Métodos estudiados en este libro suelen venir englobados habitualmente en lo que se denomina *Minería de Datos* o *Data Mining* porque suelen utilizarse con datos de gran dimensión (número $p$ de variables muy alto) y/o enorme tamaño muestral ($n$ muy grande) y, en ocasiones, con $p >> n$ lo que crea graves problemas de aplicación de las técnicas estadísticas tradicionales.

Una cuestión de interés que nos gustaría resaltar es que, en contra de lo que comúnmente se cree, este tipo de técnicas suelen ser poco robustas, es decir, suelen ser sensibles a la presencia de datos anómalos en la muestra.

También nos gustaría mencionar la referencia habitual de este tipo de técnicas, Hastie et al. (2009), aunque los objetivos de este libro y el que está leyendo son distintos.