INGENIEROS INDUSTRIALES
COLEGIO OFICIAL PRINCIPADO DE ASTURIAS

OFICINA Acelera pyme

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

UNIÓN EUROPEA

Fondo Europeo de Desarrollo Regional

"Una manera de hacer Europa"

Webinar: Data Science, el poder de los datos como elemento transformador

OFICINA Acelera pyme

INGENIEROS INDUSTRIALES
COLEGIO OFICIAL PRINCIPADO DE ASTURIAS

Duración	2 años (hasta septiembre de 2023)
Objetivo	Ir de la mano de la pyme y autónomos para ayudarles en su transformación digital.
Beneficiarios	Pymes y autónomos. Multisectorial.

Líneas de actuación

de la Oficina de transformación digital "Acelera Pyme"

Gratuito y acceso libre

Objetivo: TRANSFORMACIÓN DIGITAL DE TU EMPRESA

FORO DE TRANSFORMACIÓN DIGITAL

VÍDEO PÍLDORAS TECNOLÓGICAS

VISITAS A EMPRESAS Y HABILITADORES TECNOLÓGICOS

SESIONES DE EMPRENDIMIENTO DIGITAL

SERVICIO DE ASESORAMIENTO Y SOPORTE DIGITAL

KIT DIGITAL

JORNADAS DIVULGATIVAS EN TRANSFORMACIÓN DIGITAL

Puedes participar en todas las acciones a través de la web WWW.OTDASTURIAS.ES

Oficina de Transformación Digital “Acelera Pyme”

OFICINA Acelera pyme

INGENIEROS INDUSTRIALES PRINCIPADO DE ASTURIAS

GOBIERNO DE ESPAÑA

VICEPRESIDENCIA PRIMERA DEL GOBIERNO

MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

UNIÓN EUROPEA

Fondo Europeo de Desarrollo Regional

“Una manera de hacer Europa”

Sede del COIIAS (Oviedo)

Página web

www.otdasturias.es

RRSS

LinkedIn/Twitter/Fb/Instragram @coiias

Correo electrónico

otd@coiias.es

Suscribirse al boletín

OFICINA Acelera pyme

INGENIEROS INDUSTRIALES
PRINCIPADO DE ASTURIAS

OFICINA Acelera pyme

INGENIEROS INDUSTRIALES PRINCIPADO DE ASTURIAS

Andrés Esteban

- Graduado en Ingeniería Mecánica y Máster en Data Science e Inteligencia Artificial de Mioti Tech & Business School

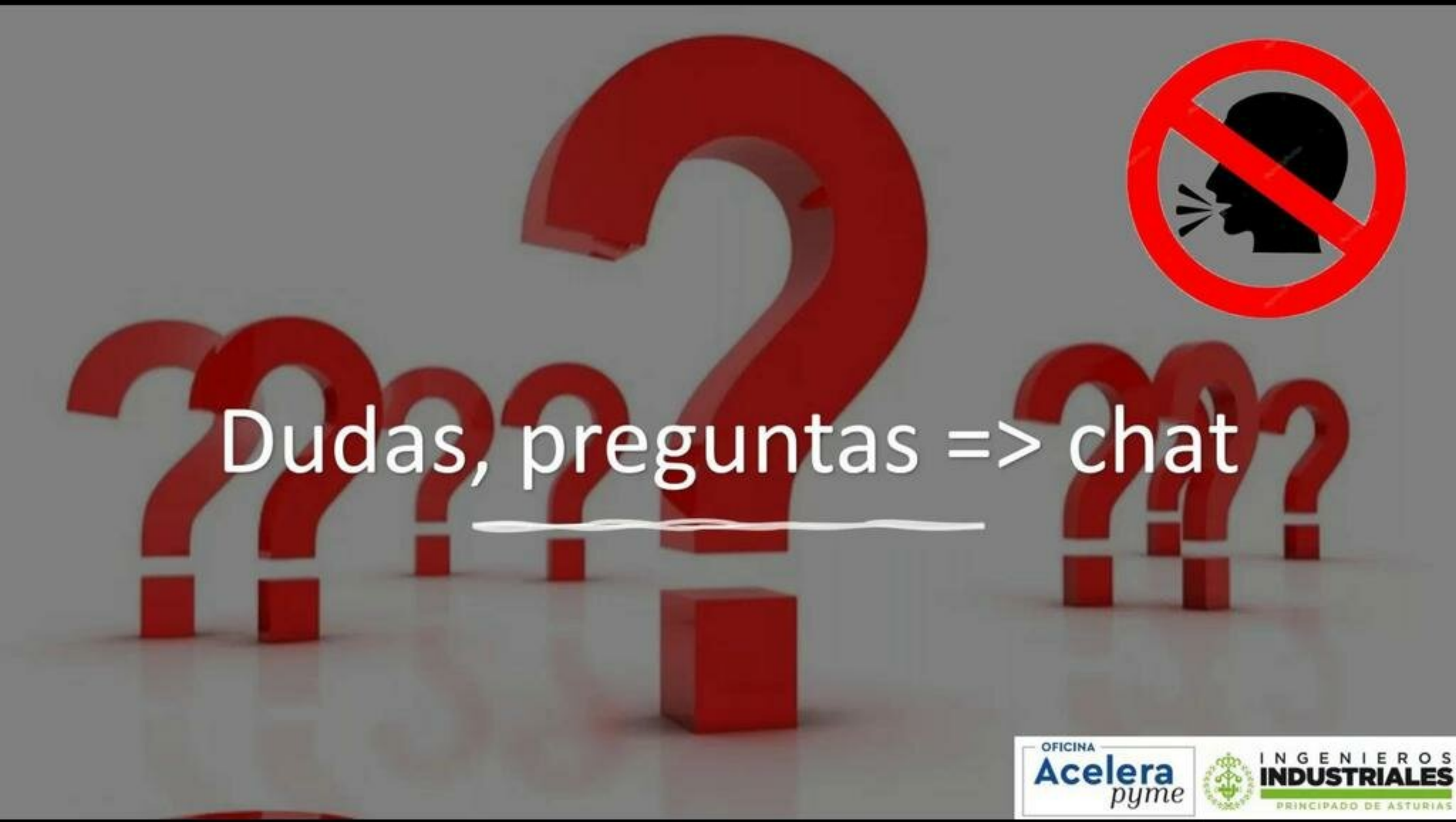
- En los últimos años ha trabajado en múltiples proyectos relacionados con el IoT y la Inteligencia Artificial en el ámbito del sector de la salud

- Hace un tiempo, ha fundado, juntos con otros dos socios, su propia Startup “Aritium”

Webinar: Data Science, el poder de los datos como elemento transformador

- Introducción a Data Science: historia, etapas de un proyecto, herramientas habituales, ejemplos reales

- Casos prácticos de aplicación


Dudas, preguntas => chat

INGENIEROS INDUSTRIALES
COLEGIO OFICIAL PRINCIPADO DE ASTURIAS

Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"

¡Gracias por Vuestra Atención!

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
UNIÓN EUROPEA

OFICINA Acelera pyme

Oficina de Transformación Digital **"Acelera Pyme"** del COIIAS

OFICINA Acelera pyme
INGENIEROS INDUSTRIALES PRINCIPADO DE ASTURIAS

Data Science & Machine Learning

aritium
the AIoT company

INGENIEROS INDUSTRIALES
PRINCIPADO DE ASTURIAS

Andrés Esteban
Co-founder & CAO
andres.esteban@aritium.com

Agenda

1. Introducción
2. Data Science
3. Machine Learning
4. Fases de un proyecto de Data Science
5. Caso práctico

1- Introducción

1.1 El poder de los datos

¿Qué porcentaje de datos utilizáis en vuestros trabajos?

1% || 2-5% || 5-10% || 10-50% || 50-75% || 75-100%

¡¡Se estima que las empresas utilizan entre el 1% y el 10% de los datos que tienen disponibles!!

1- Introducción

1.2 Olas de internet

Primera ola

Fundación de internet

IBM

Apple

hp

Sprint

Microsoft

CISCO

AMERICA Online

PC

INTERNET

1985

Segunda ola

Creados a partir de internet

amazon

Google

twitter

facebook.

ebay

PayPal

CLOUD

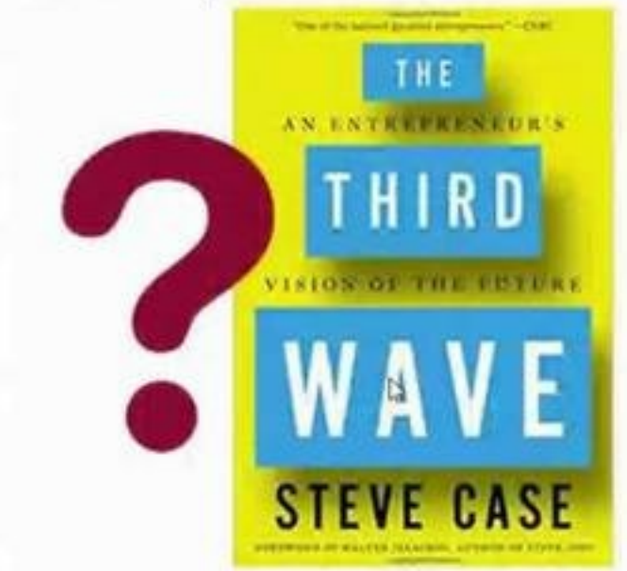
MOBILE

APP

1999

Tercera ola

La era del poder de los datos



Data Science - IA

IOT

BLOCKCHAIN

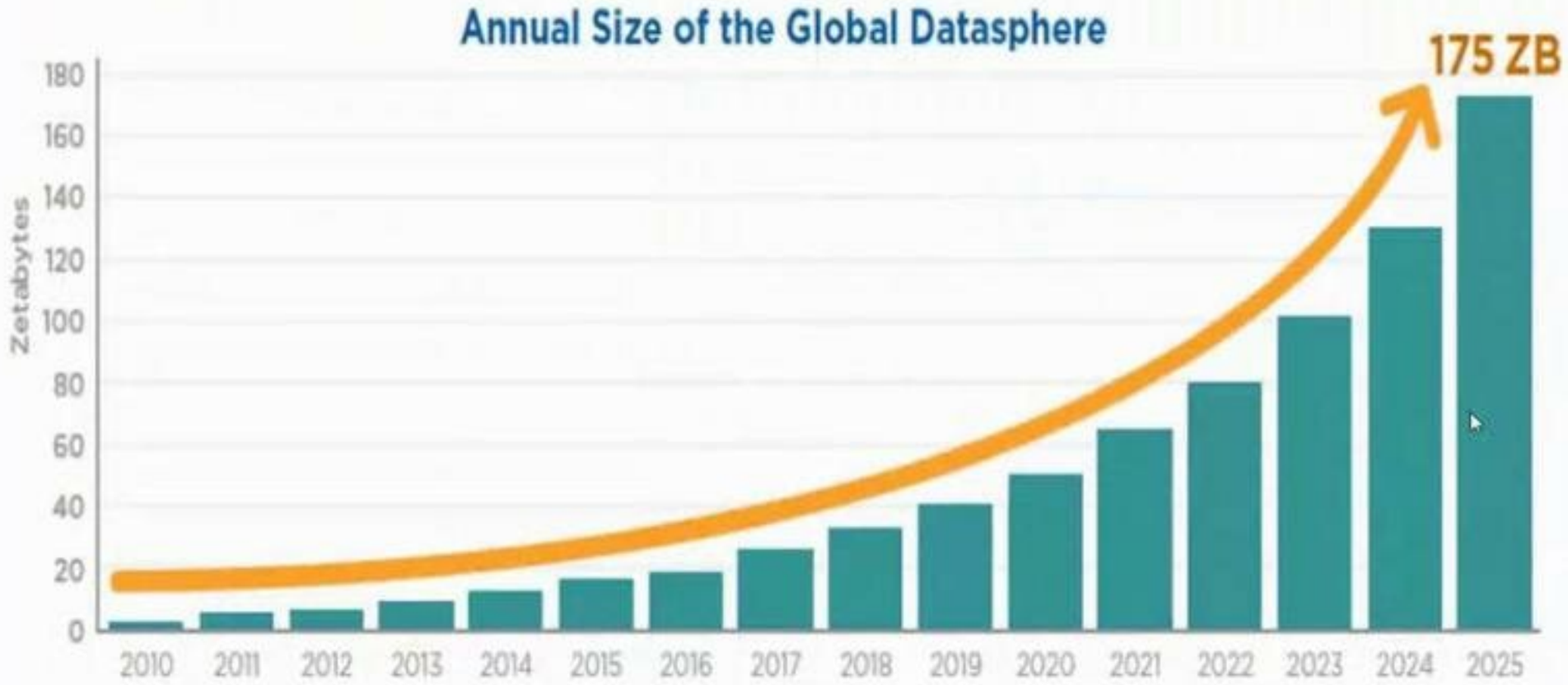
2017

1- Introducción

1.3 Evolución sensórica

1- Introducción

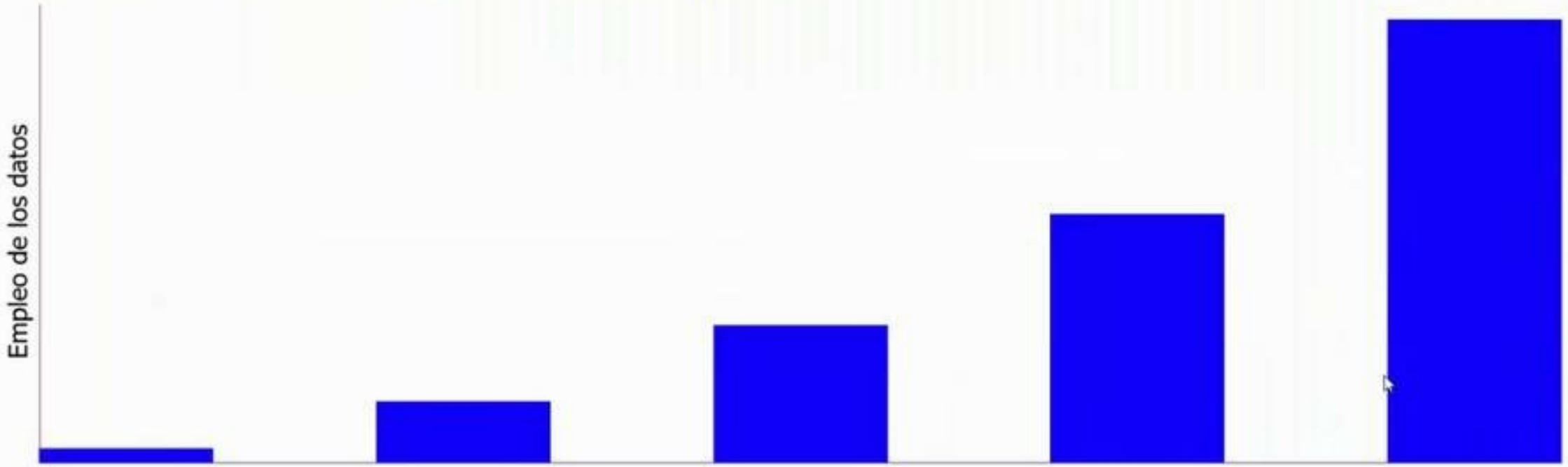
1.4 Datos generados



Se estima que el **90% de los datos** se han generado en los últimos 5 años y desde 2010 a 2020 se **multiplicó x50** el número de datos disponibles

1- Introducción

1.5 Uso de los datos por las compañías



Nivel 1
Data Resistant

Son reticentes al cambio. Creen que las cosas las van bien sin necesidad de evolucionar y piensan que tampoco se puede obtener mucho valor de la información.

Nivel 2
Data Aware

Son conscientes del valor de los datos pero están centradas en la recolección de los mismos, no los tratan correctamente y no saben cómo extraer su valor.

Nivel 3
Data Guided

Analizan lo que les dicen los datos y extraen conclusiones. Aprenden de sus errores y mejoran procesos.

Nivel 4
Data Savvy

Son compañías que aprenden de los datos, prestan atención a los datos de forma estratégica. Se centran en descubrir el porqué de estos datos. Son capaces de extraer insights.

Nivel 5
Data Driven

No solo analizan datos y extraen insights, son capaces de contestar a la pregunta: ¿qué viene después?

2-Data Science

2.1 ¿Qué es el Data Science?

Tengo un montón de datos, pero...

¿Qué hago con ellos?
¿Cómo actúo frente a los datos?
¿Cómo aprendo de los datos y genero conocimiento e inteligencia?

2-Data Science

2.1 ¿Qué es el Data Science?

¿Qué es el Data Science para vosotros?

2-Data Science

2.1 ¿Qué es el Data Science?

¿Qué es el Data Science para vosotros?

La **ciencia de datos** es un **campo interdisciplinario** que involucra **métodos científicos, procesos y sistemas** para **extraer conocimiento** o un **mejor entendimiento de datos** en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la **estadística**, la **minería de datos,** el **aprendizaje automático** y la **analítica predictiva**.

2-Data Science

2.2 Explosión del Data Science

¿Por qué ahora el auge del Data Science?

2-Data Science

2.2 Explosión del Data Science

3.Mayor desarrollo de plataformas y librerías open source

GitHub

TensorFlow

scikit learn

kaggle

Keras

PYTORCH

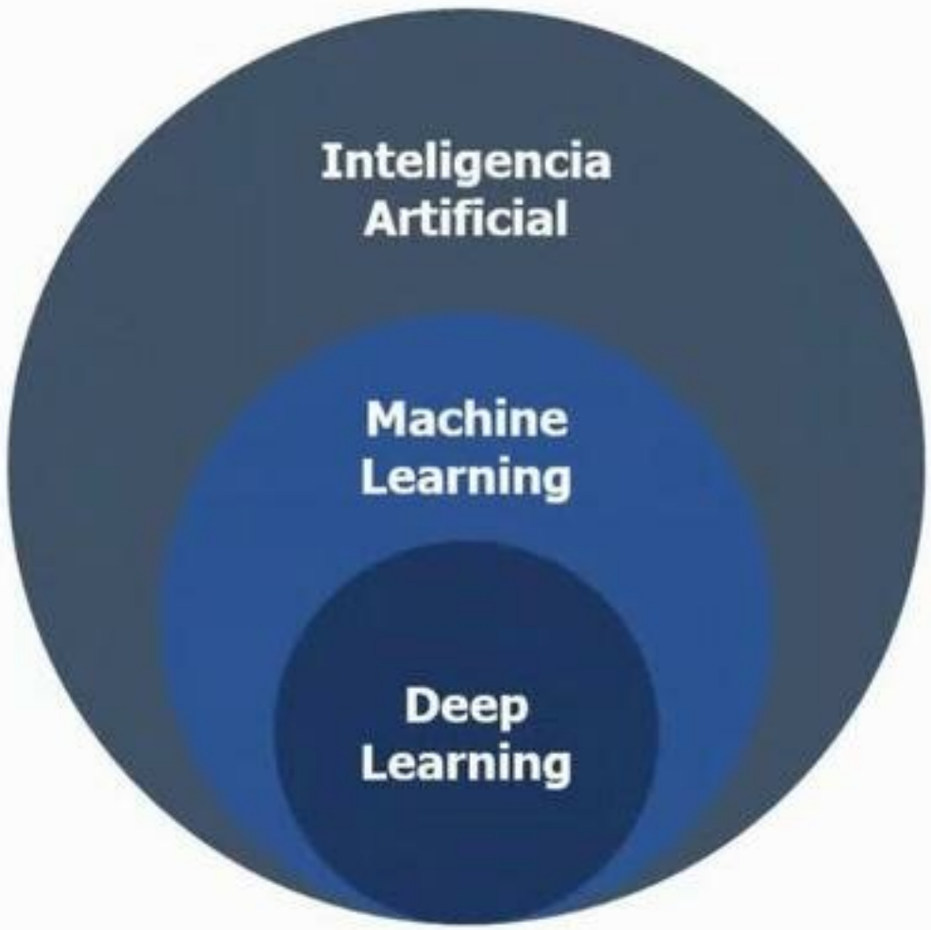
arXiv.org

theano

gensim

3-Machine Learning

3.1 IA, Machine Learning & Deep Learning



Inteligencia Artificial: Capacidad de dotar a las máquinas de ciertas habilidades propias de lo inteligencia humana, es decir, reproducen el comportamiento humano pero sin conciencia y sin aprender de ellos. Ejemplo: máquina capaz de jugar al ajedrez.

Machine Learning: disciplina dentro de la IA que, a través de algoritmos, dota a las máquinas de capacidad de aprendizaje a través de la búsqueda de patrones en conjuntos de datos. Ejemplo: la máquina aprende a jugar cada vez mejor al ajedrez

Deep Learning: disciplina dentro del Machine Learning que, a partir de una gran cantidad de datos y tras numerosas capas de procesamiento con algoritmos (redes neuronales), consigue que un ordenado aprenda por cuenta propia y realizando tareas similares a las de los seres humanos.

3-Machine Learning

3.2 Machine Learning vs Deep Learning

Machine Learning

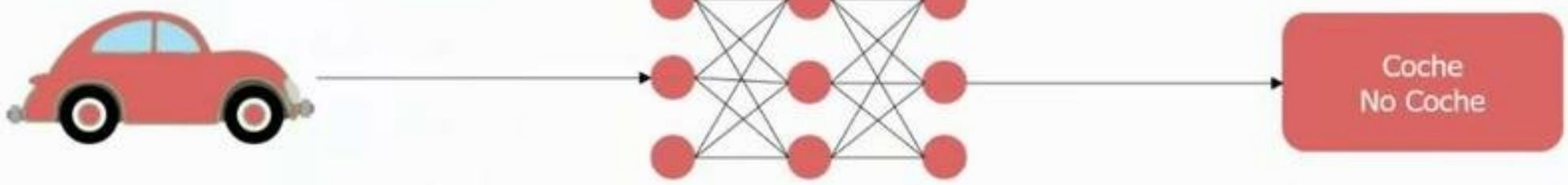
Input

Extracción de características

Algoritmo de clasificación

Output

Deep Learning



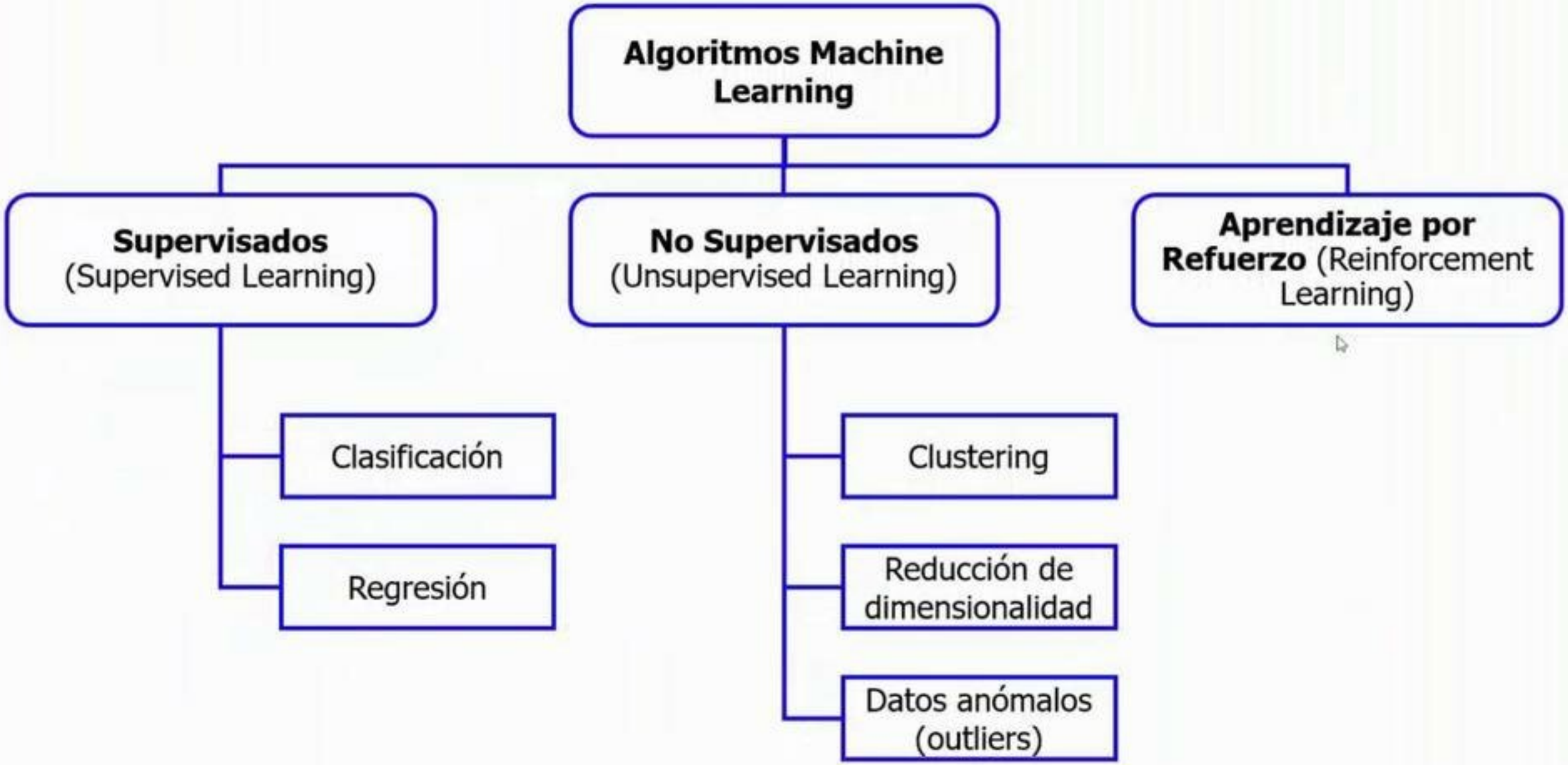
Inputs

Red neuronal
Extracción de características + clasificación

Output

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning



3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

¿Cuántas personas van a ver mi escaparate?

Queremos identificar tipos de usuarios en una red social

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

¿Cuántas personas van a ver mi escaparate?

Supervisado. Regresión

Queremos identificar tipos de usuarios en una red social

No supervisado. Clustering

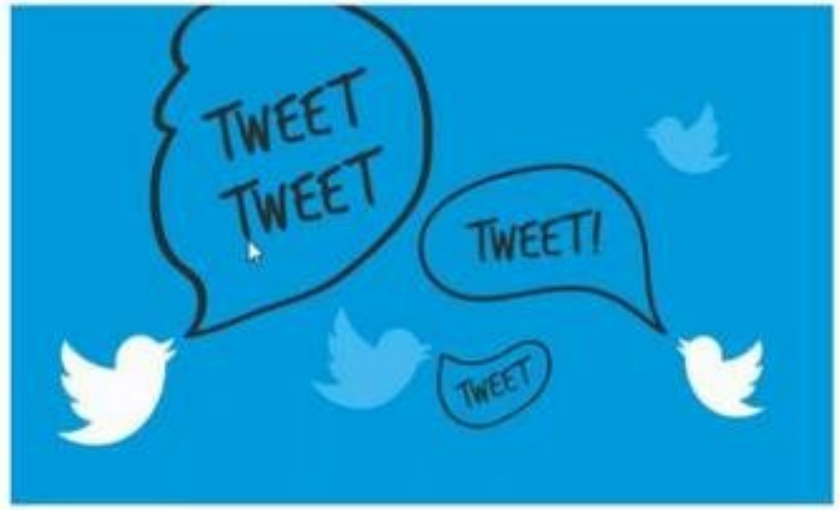
3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Vamos a hacer un gráfico en 2D para comparar el posicionamiento de las marcas de bebidas refrescantes las variables que tenemos son el precio, ventas en promoción, gasto en TV, conocimiento de marca y gasto en publicidad en internet.

No supervisado. Reducción de Dimensionalidad

Queremos detectar tweets en los que consumidores están hablando "mal" de nuestros productos



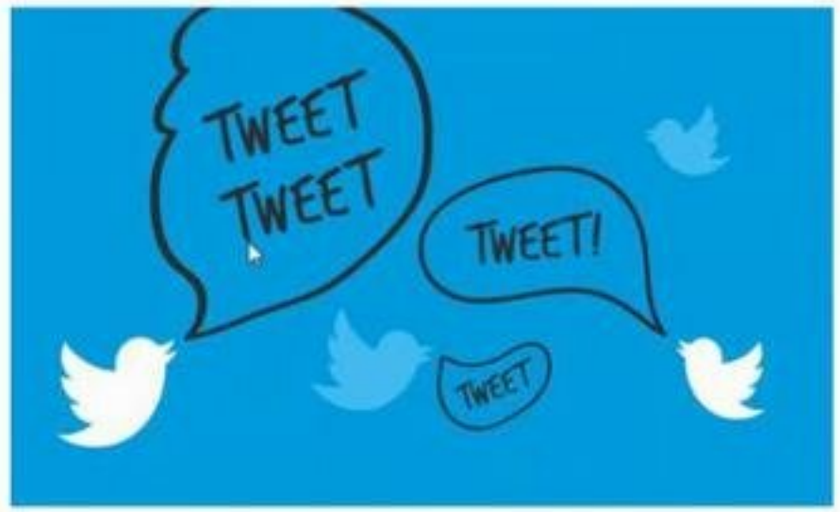
3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Vamos a hacer un gráfico en 2D para comparar el posicionamiento de las marcas de bebidas refrescantes las variables que tenemos son el precio, ventas en promoción, gasto en TV, conocimiento de marca y gasto en publicidad en internet.

No supervisado. Reducción de Dimensionalidad

Queremos detectar tweets en los que consumidores están hablando "mal" de nuestros productos



Supervisado. Clasificación

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

¿Cuándo tengo que enviar a un técnico para realizar mantenimiento predictivo?

Supervisado. Regresión

Queremos analizar qué diferentes tipos de conductores existen en función del tipo de conducción que realizan Para eso recogemos datos de conducción en tiempo real durante 2 semanas de conducción (velocidad, aceleración, cómo frenan, ….)

No supervisado. Clustering

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Queremos un robot que aprenda cómo conseguir limpiar una habitación en el menor tiempo posible

Queremos estimar la evolución de nuestras ventas en los próximos 5 años

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Queremos un robot que aprenda cómo conseguir limpiar una habitación en el menor tiempo posible

Aprendizaje por refuerzo

Queremos estimar la evolución de nuestras ventas en los próximos 5 años

Supervisado. Regresión

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Queremos segmentar nuestros consumidores utilizando la información que disponemos de ellos en nuestra página web (tiempo de conexión, número de veces, páginas que visitan, ….)

No supervisado. Clustering

Quiero detectar qué clientes me van a dejar de comprar el mes que viene

Supervisado. Clasificación

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Quiero detectar cuántos clientes me van a dejar de comprar el mes que viene

Supervisado. Regresión

Queremos detectar si existe fraude en una transacción realizada con una tarjeta de crédito

Supervisado. Clasificación

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

Queremos entrenar un algoritmo que aprenda a jugar al póker

Aprendizaje por Refuerzo

Modelo que me permita detectar cuántas personas van a estar enfermas de gripe el próximo año

Supervisión. Regresión

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

¿Qué preferimos…?

Modelo supervisado

Modelo **NO** supervisado

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

¿Qué preferimos…?

Modelo supervisado

Modelo **NO** supervisado

Aprender

Descubrir

3-Machine Learning

3.2 Clasificación algoritmos Machine Learning

¿Por qué tenemos que aprender sin un profesor?

Si nuestro objetivo es crear sistemas inteligentes que puedan realizar un conjunto de tareas diferentes, ¿por qué no enseñamos directamente las tareas mediante un modelo supervisado o con aprendizaje reforzado?

1. Las etiquetas / recompensas pueden ser **difíciles de obtener o definir** y estas contienen **menos información que el dato que tenemos de entrada**.
2. El **aprendizaje no supervisado** se asemeja en mayor medida a la forma en la que **aprendemos los humanos.**
3. Buscamos conocimiento que nos permita **generalizar y responder a nuevas tareas y situaciones**.
4. El **aprendizaje no supervisado nos da información sobre las relaciones de los datos** y muchos de ellos nos pueden sorprender. ¿no vamos a utilizarlo?

3-Machine Learning

3.3 Principales problemas Machine Learning

Problema de sobreajuste (overfiting)

Aprender VS Memorizar

Calidad de los datos

"La calidad de un modelo está limitada por la calidad de los datos utilizados para entrenarlo"

Principio GIGO

Agenda

1. Introducción
2. Data Science
3. Machine Learning
4. **Fases de un proyecto de Data Science**
5. Caso práctico

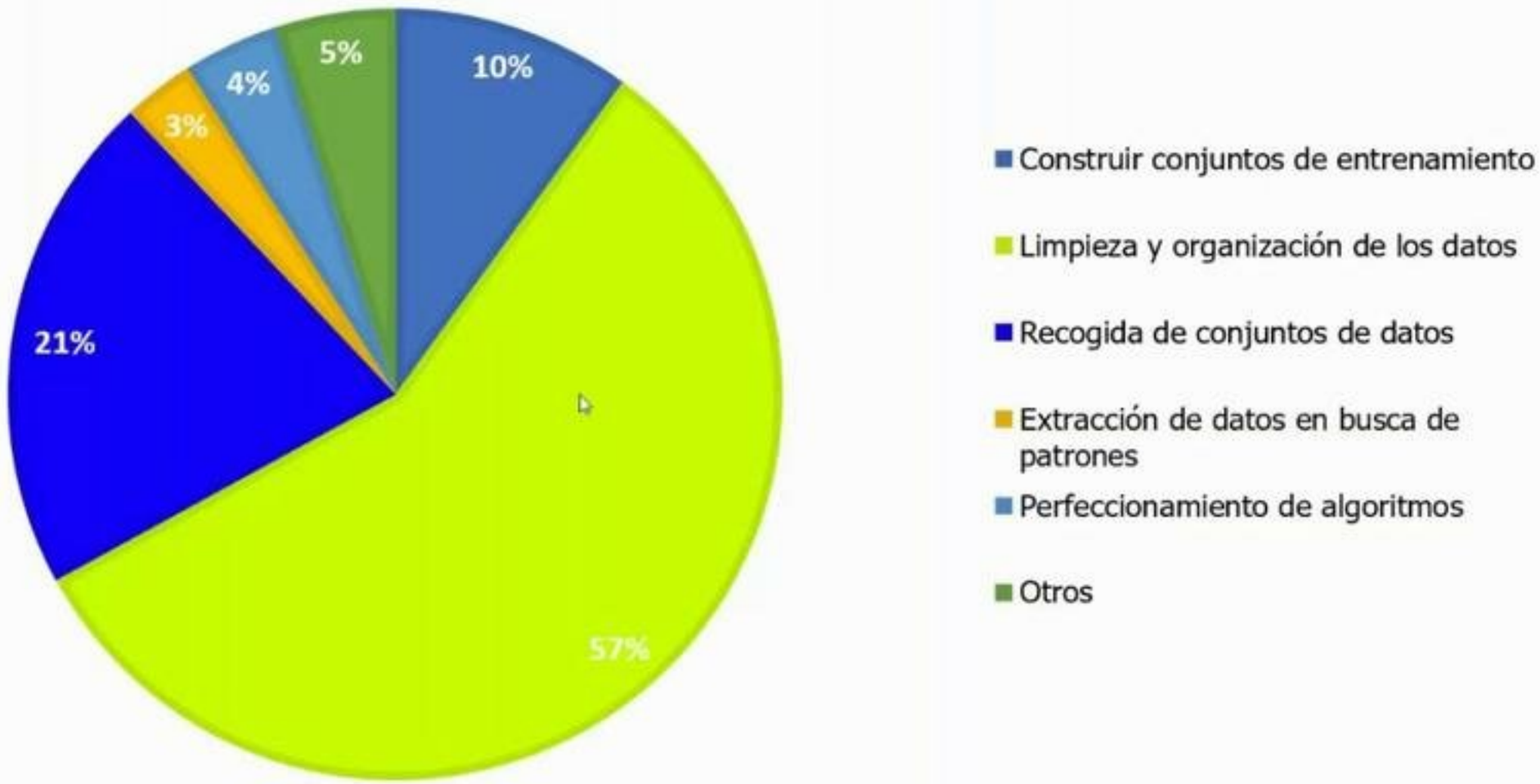
4-Fases de un proyecto de Data Science

4.1 Esquema principal

1. **Visión general** del problema
2. **Obtención** de los datos
3. **Visualizar** y analizar los datos
4. **Preparar** los datos para la modelización
5. **Selección** un algoritmo y entrenarlo
6. **Ajustar** el modelo, análisis de errores
7. **Presentar** resultados
8. **Poner** en producción el modelo

4-Fases de un proyecto de Data Science

4.2 Porcentaje de cada fase



5-Caso práctico

5.1 Fase 1: Visión general del problema

¿Por qué predecir la duración de las intervenciones?

SITUACIÓN ACTUAL

La demora de las intervenciones quirúrgicas es uno de los mayores retos a los que se enfrentan los servicios sanitarios de todo el mundo

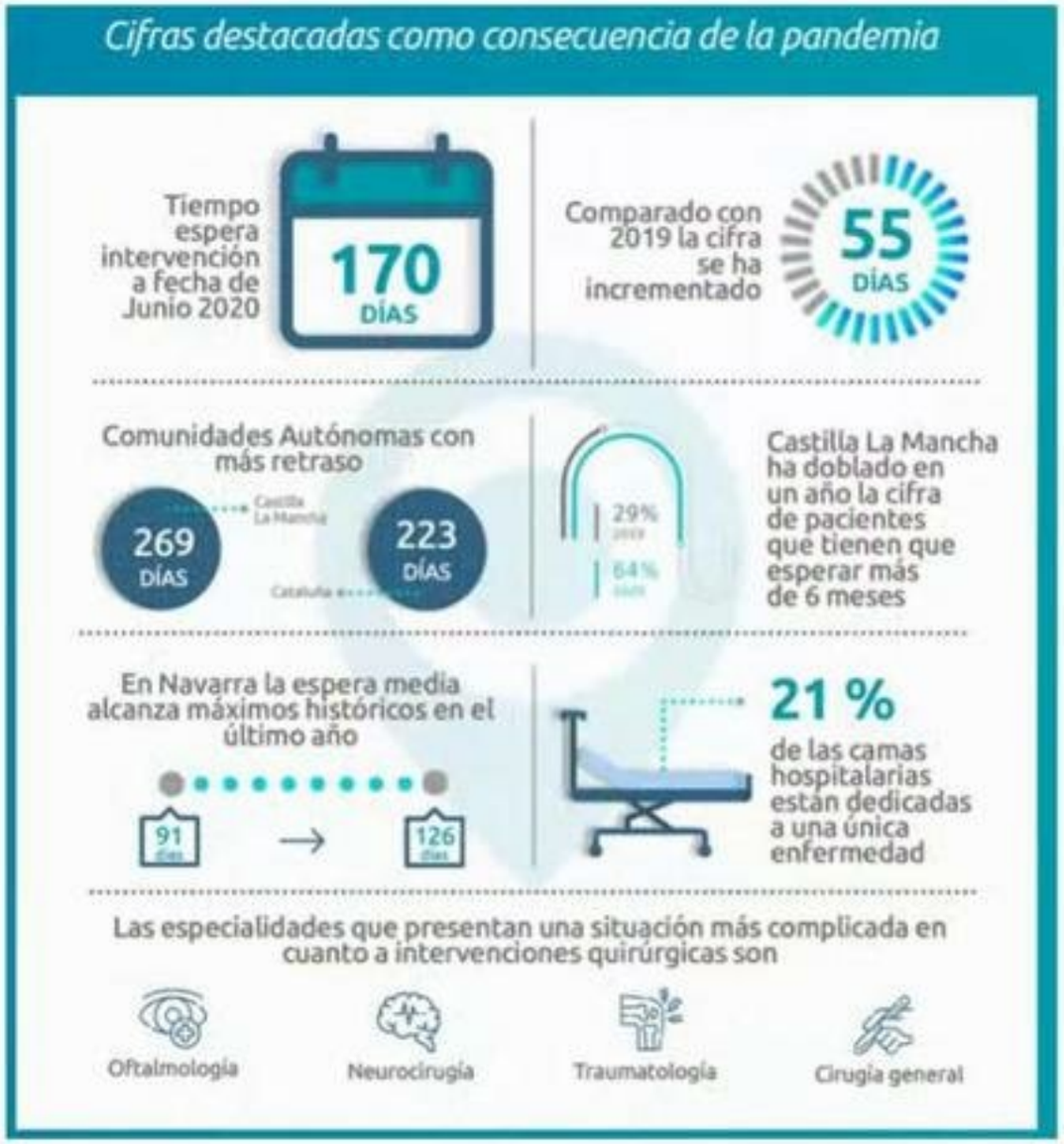
- En España se calcula que las operaciones en espera superan el millón.
- La pandemia ha agravado este problema con una demora media de 170 días.

SOLUCIONES ACTUALES

Algunas de las medidas tomadas han sido:

- Refuerzo de los quirófanos a través de conciertos con la sanidad privada.
- Derivación de pacientes a centros privados.
- Incremento de la actividad quirúrgica y aumento de la remuneración al personal.

Sin embargo, **NO ES SUFICIENTE**. Por ello, es necesario apoyarse en nuevas herramientas que aumenten el rendimiento quirúrgico **-> IoT e IA**.



5-Caso práctico

5.2 Fase 2: Obtención de los datos

Se ha empelado un dataset que contiene un histórico de intervenciones quirúrgicas desde el año 2017 hasta el 2020. En total contiene 172.722 intervenciones con 45 variables diferentes.

```
Data columns (total 45 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Unnamed: 0              172722 non-null  int64
 1   ID                      172722 non-null  int64
 2   ejecucion               172722 non-null  int64
 3   programacion            132403 non-null  float64
 4   programada              172722 non-null  int64
 5   quirofano               172722 non-null  object
 6   anestesia               172722 non-null  object
 7   medico_solicitante      159019 non-null  float64
 8   servicio_intervencion   172722 non-null  object
 9   tipo_intervencion       172721 non-null  object
 10  tipo_cirugia            140704 non-null  object
 11  procedencia             172721 non-null  object
 12  realizada               172722 non-null  object
 13  motivo_cancelacion      6743 non-null    object
 14  destino                 167487 non-null  object
 15  fecha_entrada_quirofano 172722 non-null  datetime64[ns]
 16  fecha_inicio_cirugia    158379 non-null  datetime64[ns]
 17  fecha_fin_cirugia       158382 non-null  datetime64[ns]
 18  fecha_salida_quirofano  172716 non-null  datetime64[ns]
 19  diagnostico_cie         139471 non-null  object
 20  procedimiento_cie       138634 non-null  float64
 21  ecod0                   146906 non-null  float64
 22  ecat0                   146906 non-null  object
 23  epri0                   146906 non-null  object
 24  ecod1                   124297 non-null  float64
 25  ecat1                   124297 non-null  object
 26  epri1                   124297 non-null  object
 27  ecod2                   95170 non-null   float64
 28  ecat2                   95170 non-null   object
 29  epri2                   95170 non-null   object
 30  ecod3                   62690 non-null   float64
 31  ecat3                   62690 non-null   object
 32  epri3                   62690 non-null   object
 33  ecod4                   44236 non-null   float64
 34  ecat4                   44236 non-null   object
 35  epri4                   44236 non-null   object
 36  ecod5                   25473 non-null   float64
 37  ecat5                   25473 non-null   object
 38  epri5                   25473 non-null   object
 39  ecod6                   10556 non-null   float64
 40  ecat6                   10556 non-null   object
 41  epri6                   10556 non-null   object
 42  fecha                   172722 non-null  datetime64[ns]
 43  fecha_nacimiento        172722 non-null  datetime64[ns]
 44  edad                    172722 non-null  int64
dtypes: datetime64[ns](6), float64(10), int64(5), object(24)
memory usage: 59.3+ MB
```

5-Caso práctico

5.3 Fase 3&4: Visualizar y analizar los datos y prepararlos para la modelización

PROCESAMIENTO DE LOS DATOS

Preprocesamiento del dataset inicial a partir del cual se obtienen 4 datasets diferentes con un total de **135.845 intervenciones** para evaluar diferentes modelos de clasificación

PREPROCESADO GENERAL

- **BORRADO DE VARIABLES** que no son de interés (ejemplo: intervenciones no realizadas)
- **BORRADO DE MISSING VALUES** (ejemplo: intervenciones sin procedimiento CIE)
- **IMPUTACIÓN DE MISSING VALUES** (ejemplo: tiempo de ocupación quirófano vs duración intervención)
- **FEATURE ENGINEERING** (ejemplo: profesionales que participan en cada intervención)
- **ONE HOT ENCODING** de las variables categóricas (ejemplo: anestesista, tipo de intervención)
- **AGRUPACIÓN DE VARIABLES** (ejemplo: procedimiento CIE los 2153 valores lúnicos lo agrupamos en los 20 más frecuentes)
- **DEFINICIÓN CLASES Y ETIQUETADO**
- **ANÁLISIS EXPLORATORIO DE LOS DATOS**
  - Matriz de correlación de variables
  - Histogramas
  - Análisis Estadístico Descriptivo

Preprocesamiento 1

- **6 clases según criterio experto médico**
- **5 clases según división quintiles**
- **165 variables**

Preprocesamiento 2

- **6 clases según criterio experto médico**
- **5 clases según división quintiles**
- **184 variables**

Preprocesamiento 3

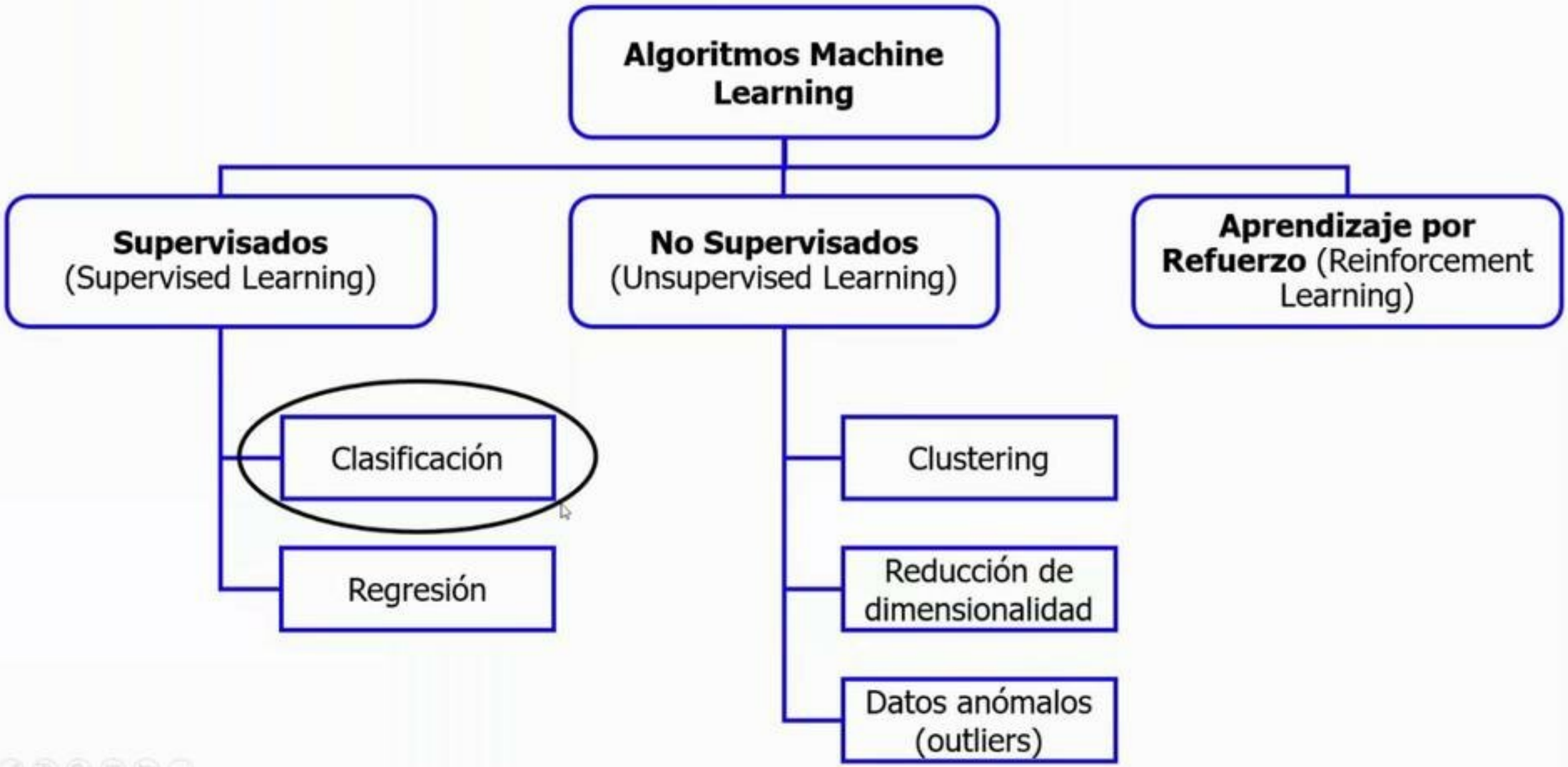
- **14 clases según criterio experto médico**
- **5 clases según división quintiles**
- **165 variables**

Preprocesamiento 4

- **14 clases según criterio experto médico**
- **5 clases según división quintiles**
- **184 variables**

5-Caso práctico

5.4 Fase 5: Selección de algoritmo y entrenamiento



5-Caso práctico

5.4 Fase 5: Selección de algoritmo y entrenamiento

Algoritmos Machine Learning

Supervisados
(Supervised Learning)

Clasificación

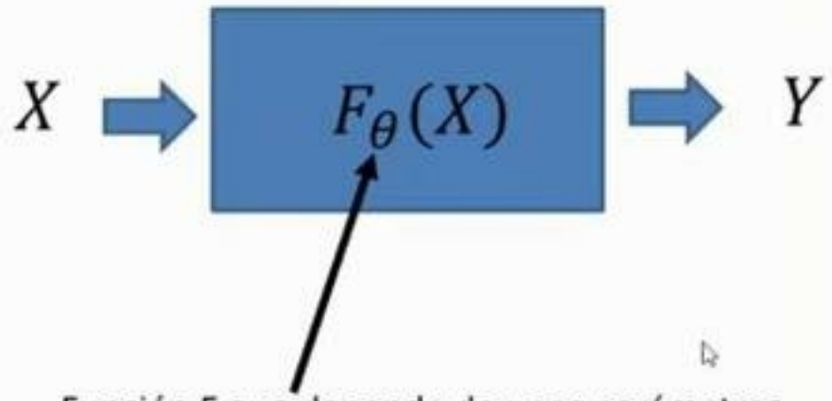
RandomForest	XGBoost	SVM	MLP (RN) 2 & 3 capas
👍 👍 Coste Computacional	👍 👍 Coste Computacional	👎 👎 Coste Computacional	👍 Coste Computacional
👍 👍 Métricas	👍 👍 Métricas	👎 Métricas	👍 Métricas
💀 **SOBREAJUSTE** 💀	💪 NO SOBREAJUSTE	💪 NO SOBREAJUSTE	💪 NO SOBREAJUSTE

5-Caso práctico

5.4 Fase 5: Selección de algoritmo y entrenamiento

CONEPTOS CLAVE - ¿Cómo funcionan los algoritmos de aprendizaje supervisado?

Aprendizaje supervisado

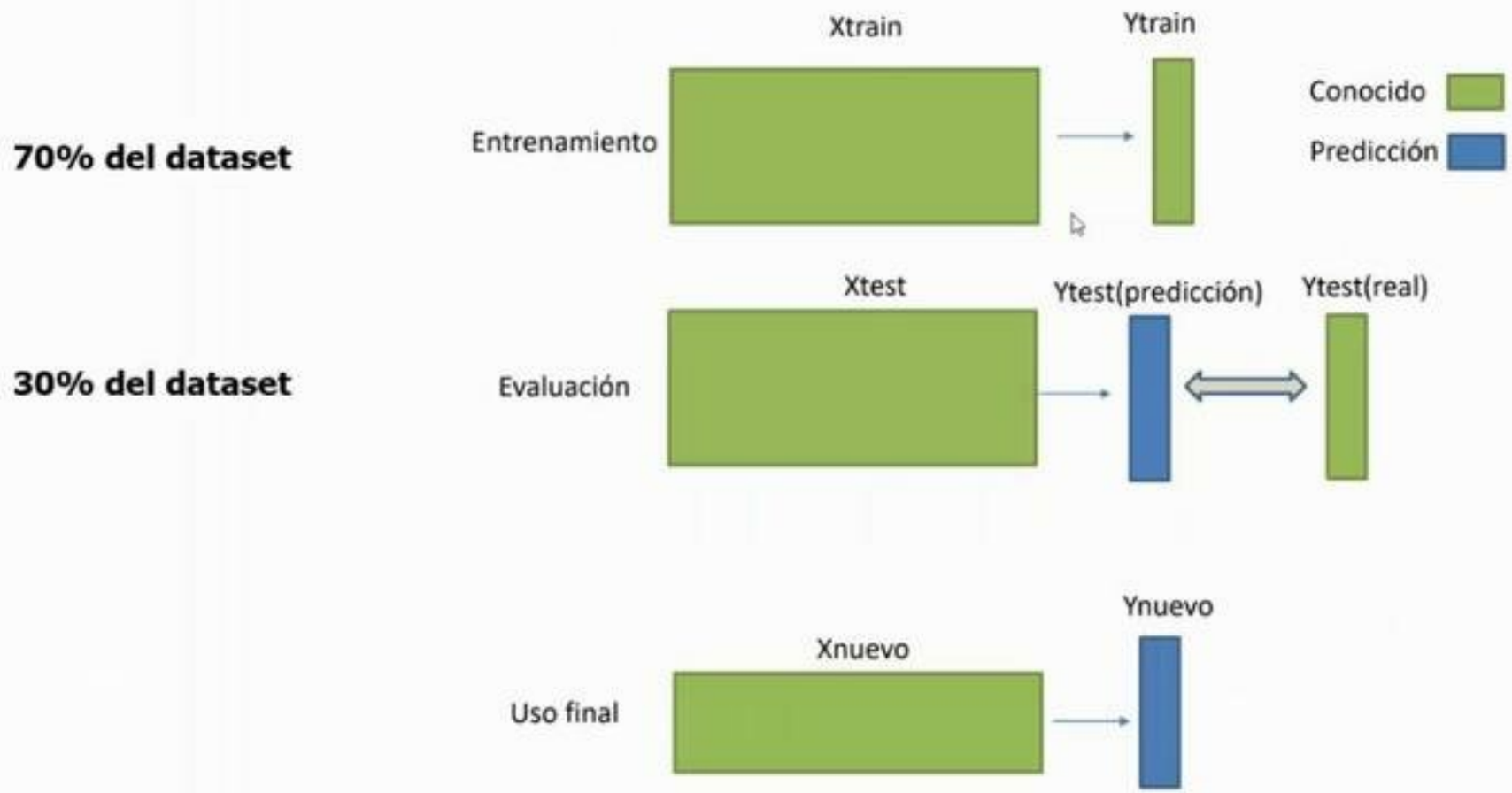


Los parámetros $\boldsymbol{\theta}$ los **aprendemos** de manera **automática** a partir de datos de entrenamiento $\boldsymbol{X_{train}}$, $\boldsymbol{Y_{train}}$

5-Caso práctico

5.4 Fase 5: Selección de algoritmo y entrenamiento

CONEPTOS CLAVE – Entrenamiento y test



5-Caso práctico

5.5 Fase 6: Ajustar el modelo, análisis de errores

CONEPTOS CLAVE – Hiperparámetros

Model	Overview	Hyperparameters
C4.5	J48 Decision Tree	c ={0.05,0.10,0.20,0.30,0.40,0.50,0.60,0.70}
NNET	3-layer Neural Network	size = {4,..,28}, decay = {0.10,0.20}
KNN	K- Nearest Neighbor	c ={2*(0,.....,7) + 1}
RF	Random Forest	mtry= { 10, 50,100, 200, 250,500, 1000}
SVM	Support Vector Machine	c ={2^{-6},. . . .,2^{10}}

5-Caso práctico

5.6 Fase 7: Presentar resultados

CONEPTOS CLAVE – Métodos de evaluación

PRECISION || RECALL || ACCURACY || F1 SCORE

Predicho

Real		SI	NO	
	SI	4	7	11
	NO	1	119	120
		5	126	

A

Precision = 4/5 = 0.8
Recall = 4/11 = 0.36
Accuracy = 123/131 = 0.94

Predicho

Real		SI	NO	
	SI	10	1	11
	NO	20	100	120
		30	101	

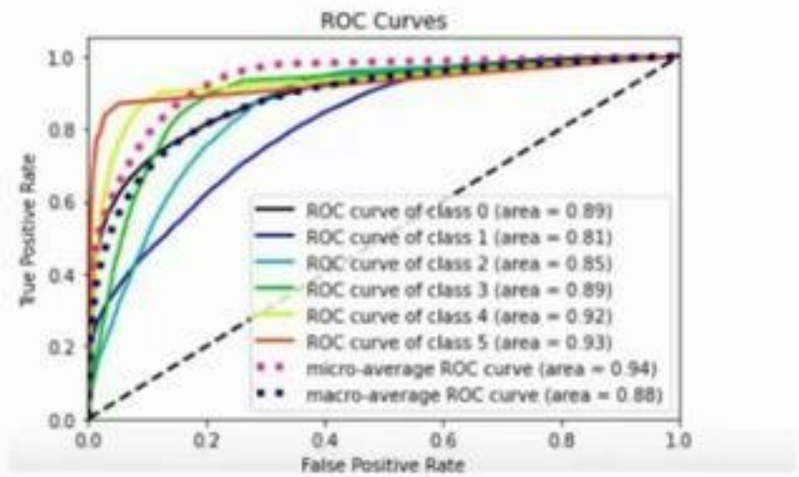
B

Precision = 10/30 = 0.3
Recall = 10/11 = 0.91
Accuracy = 110/131 = 0.84

Matriz de confusión

- **Precision** -> precisión, se puede medir la calidad del modelo en tareas de clasificación.
- **Recall** -> exhaustividad, informa sobre la cantidad de datos que el modelo es capaz de identificar.
- **F1 Score** -> combinación de Precision y Recall en un solo valor (nos importan ambos por igual)
- **Accuracy** -> exactitud, mide el porcentaje de casos que el modelo ha acertado.

CURVAS ROC



5-Caso práctico
5.6 Fase 7: Presentar resultados

ELECCIÓN DEL MODELO FINAL - XGBoost

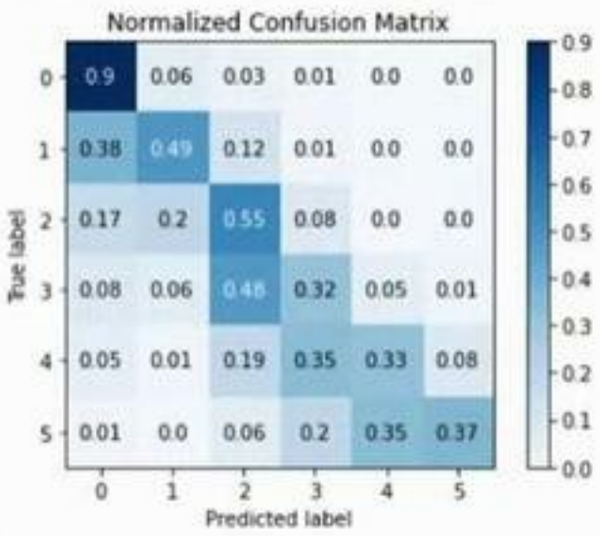
★ Clasificador con **mejores resultados** y **menor coste computacional** con los siguientes hiperparámetros

colsample_bytree= 0.7545474901621302
learning_rate= 0.12336180394137353
max_depth= 14
n_estimators= 24
subsample= 0.6028265220878869

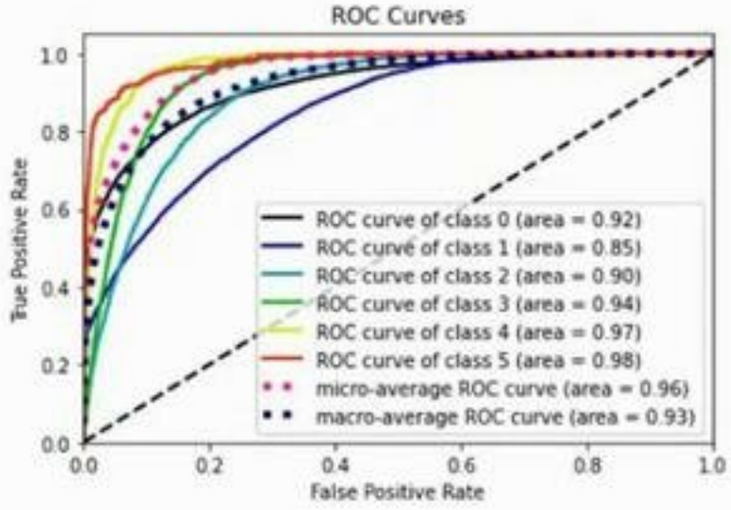
XGBoost

👍 👍 Coste Computacional
👍 👍 Métricas
💪 NO SOBREAJUSTE

★ **Dataset #2** → Clases según criterio médico (**6 clases**), **184 variables** y **138.845 instancias**




```
Accuracy score TEST MEDICO: 0.7239534769593169
Precision score TEST MEDICO: 0.71413526479939
Recall score TEST MEDICO: 0.7239534769593169
F1 score TEST MEDICO: 0.7147251225803479
```



5-Caso práctico

5.7 Fase 8: Poner en producción el modelo


gracias