UAX Universidad Alfonso X el Sabio

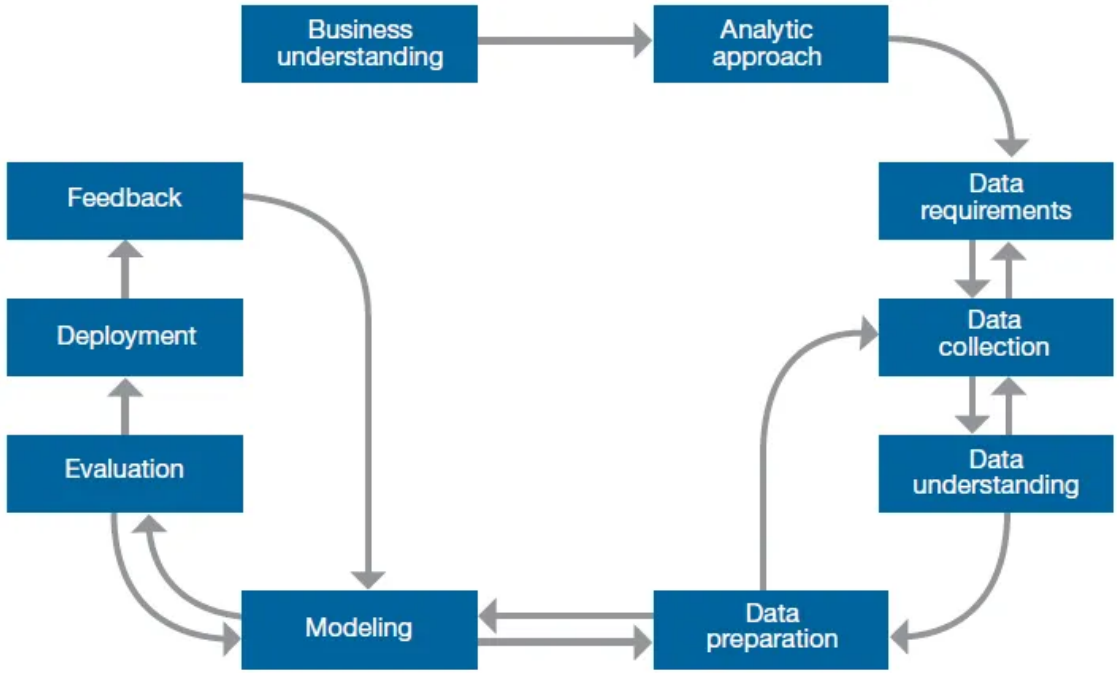
CIENCIA DE DATOS

GRADO EN BIOMEDICINA

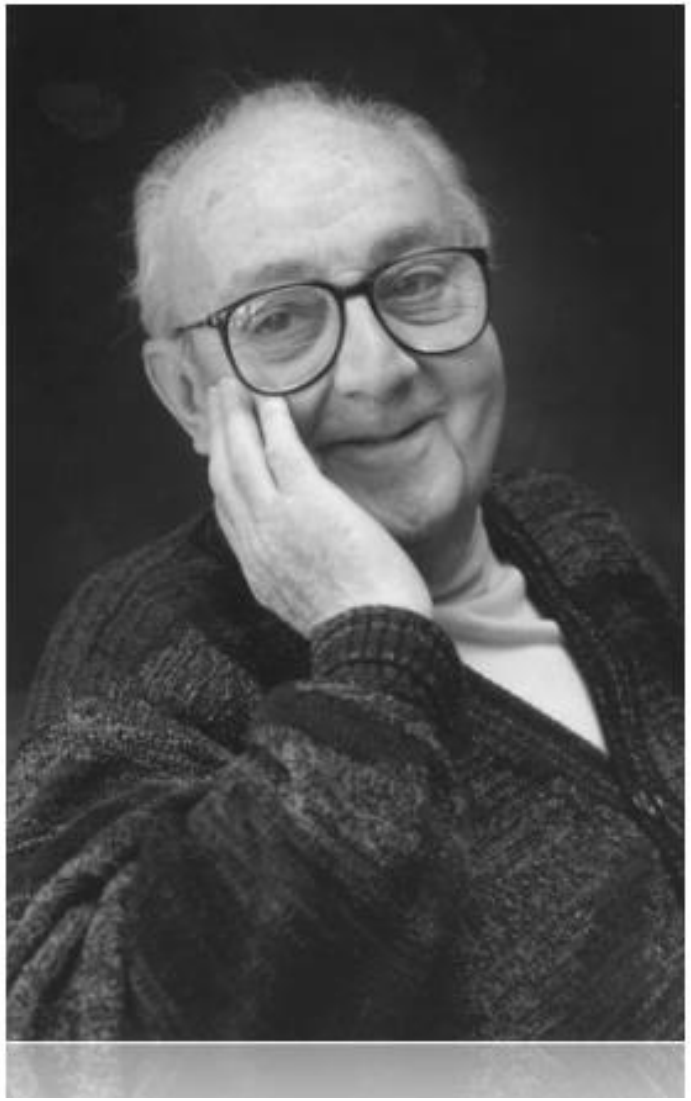
DR. ARIEL CARIAGA-MARTÍNEZ

¿Qué es el modelado de datos?

El modelado de datos es el proceso de estructurar y organizar los datos para su uso efectivo en análisis y predicciones. En ciencia de datos, el modelado no solo implica la organización física de los datos, sino también la posibilidad de desarrollar posibles abstracciones (MATEMÁTICA), relaciones, reglas y estructuras entre ellos.



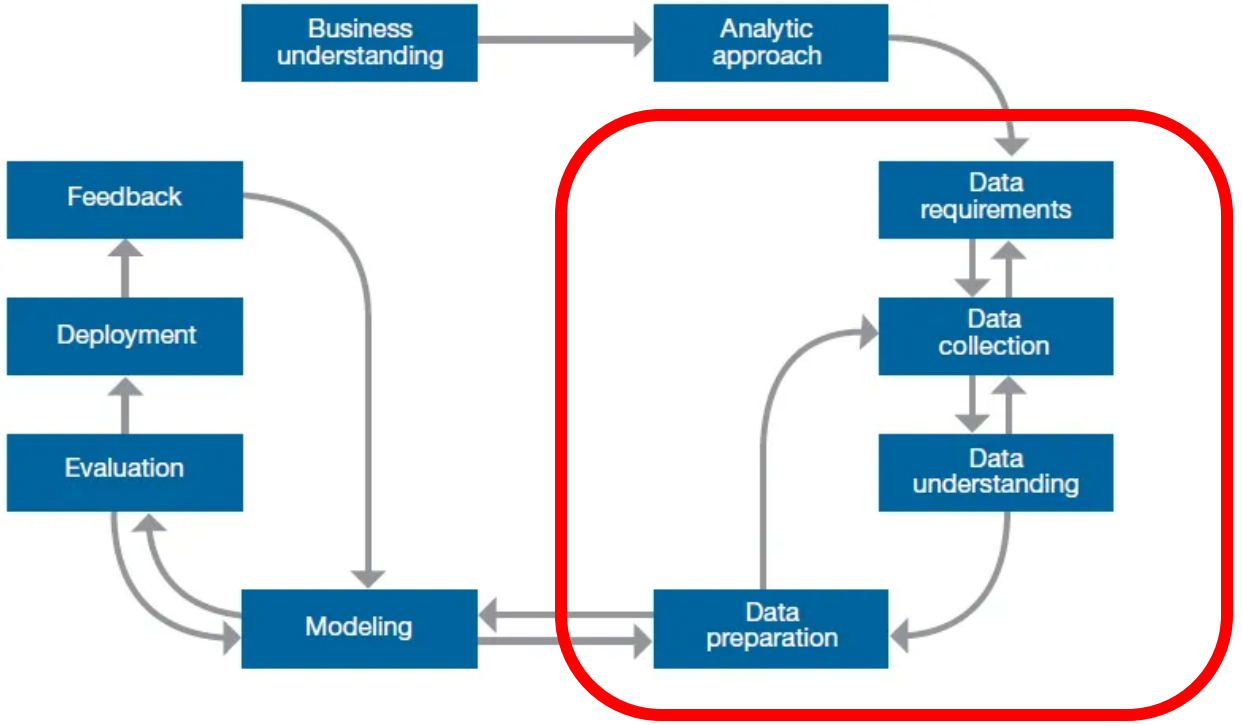
https://medium.com/@applying.pe/10-etapas-para-la-ciencia-de-datos-b4689181d0a2



"Essentially, all models are wrong, but some are useful."

– *George E. P. Box*

Pasos previos



- - Formato que facilite su análisis y manipulación.
- - Inconsistencias, duplicaciones y datos → EDA EN PROFUNDIDAD.
- - Conclusión: tomar decisiones fundamentadas sobre qué datos usar, cómo deben ser tratados y cómo deben ser interpretados, reduciendo así errores potenciales en el análisis final.

¿Qué modelo elegir?

PREGUNTA + DATOS (LIMPIOS) DISPONIBLES + USO (CASOS POSIBLES)

- **CLASIFICACIÓN**
- **REGRESIÓN**
- **CLÚSTERES**
- **SERIES TEMPORALES**
- **ETC.**

Preparación de los datos para el modelado.

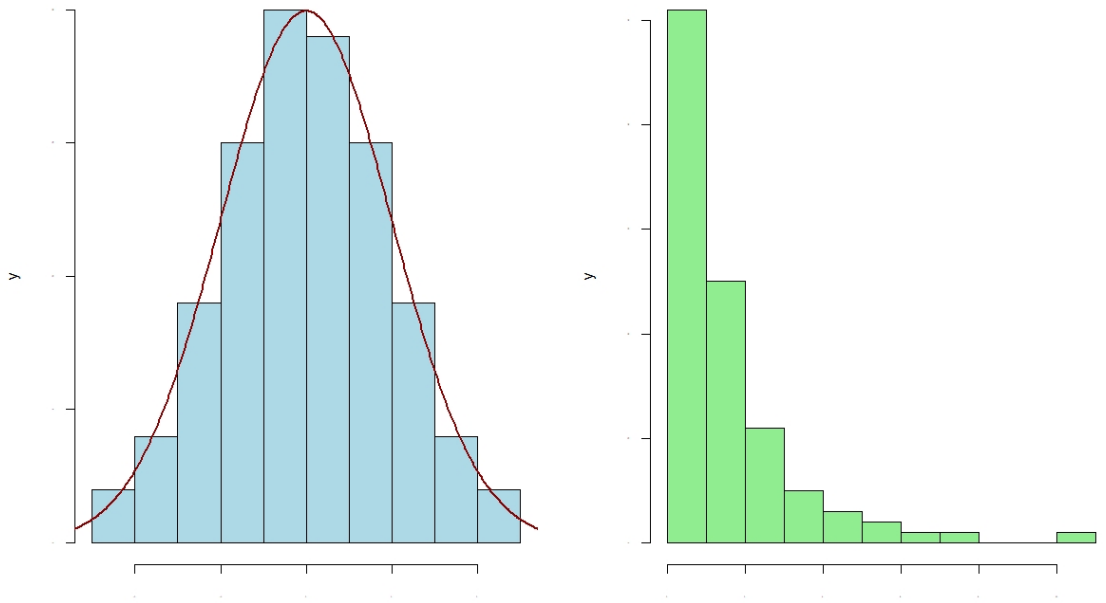
La normalización, escalado y transformación de los datos son pasos cruciales para asegurar que el análisis sea coherente.

- La normalización ajusta los datos para que estén dentro de un mismo rango, por ejemplo, entre 0 y 1, eliminando diferencias de magnitud entre las variables. **X_normalizado = (X - X_min) / (X_max - X_min)**

- El escalado o estandarización busca ajustar las variables a distribuciones (típicamente media = 0 y var = 1): Z-score

- Transformación: incluye técnicas como la conversión/normalización logarítmica, normalización decimal.

- Normalización robusta usando percentiles (25-75) para evitar la influencia de outliers.



“Ingeniería de variables”

Transformar datos en nuevas variables para mejorar el rendimiento de los modelos.

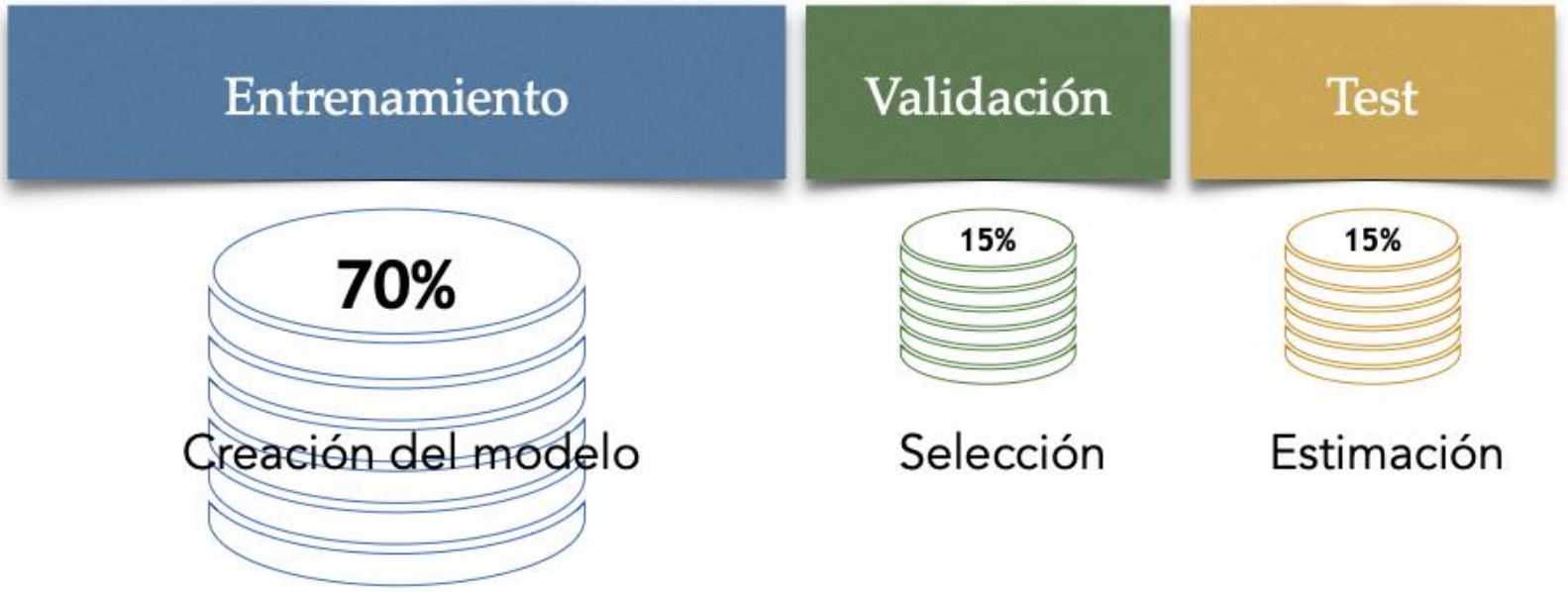
- **- Creación de variables derivadas (de fechas a edades).**
- **- Relaciones entre variables (sumas, multiplicaciones, ratios).**
- **- Binarizaciones y categorizaciones.**
- **- Selección automática vs estadística**

“Ingeniería de variables”: más opciones

- - Agregación de datos (series temporales): sumas, promedios incluso desviaciones.
- - Interacciones entre variables (modeladas “aparte”).
- **- Reducir la dimensionalidad.**
- **- Creatividad…matemática… + contexto**

Dividiendo los datos para la puesta a punto

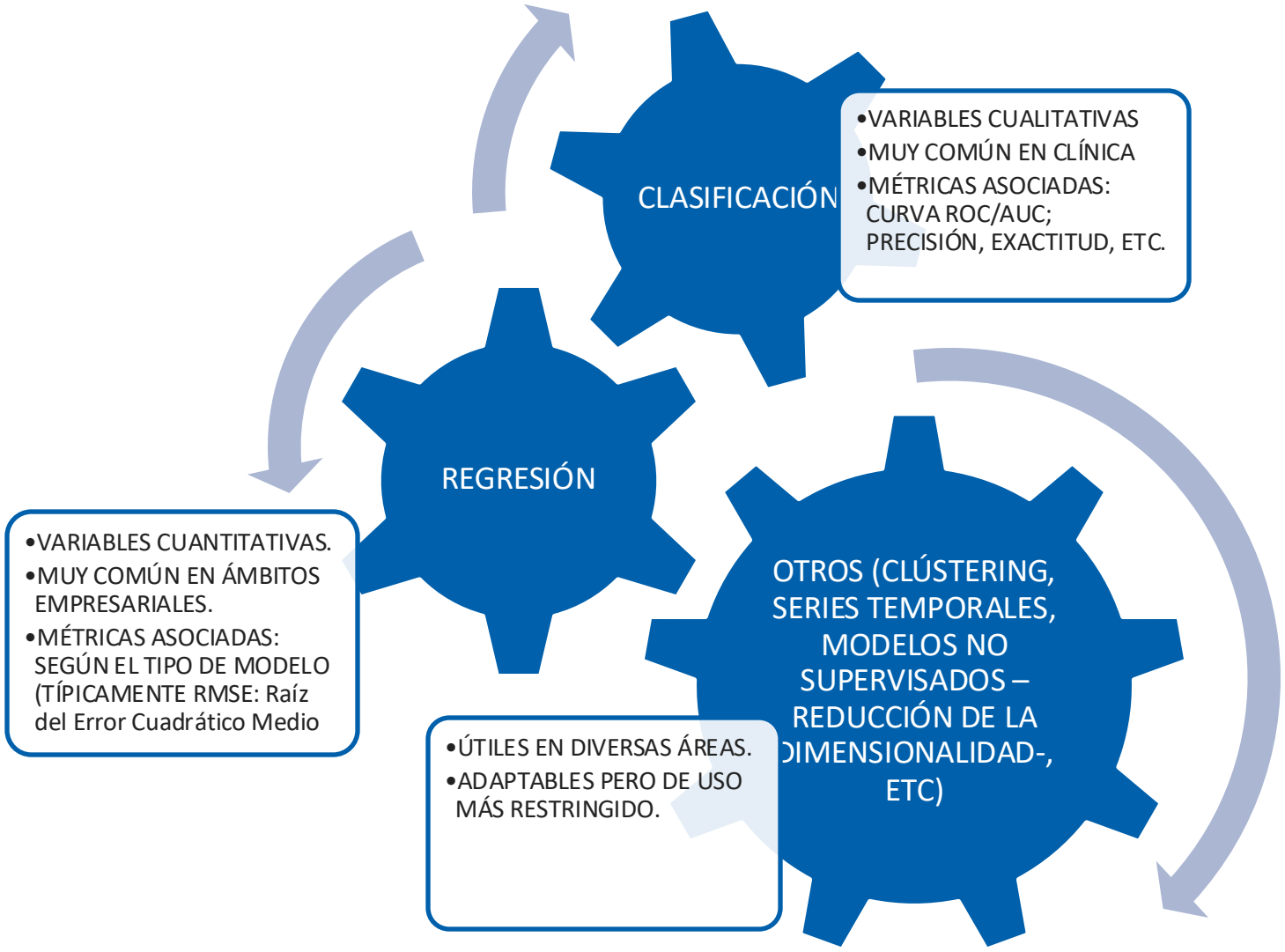
- Entrenamiento: se utiliza para ajustar el modelo.
- Validación: ayuda a optimizar los hiperparámetros del modelo + generalización.
- Prueba: evalúa el rendimiento final del modelo en datos no utilizados previamente.



80/20: sin validación

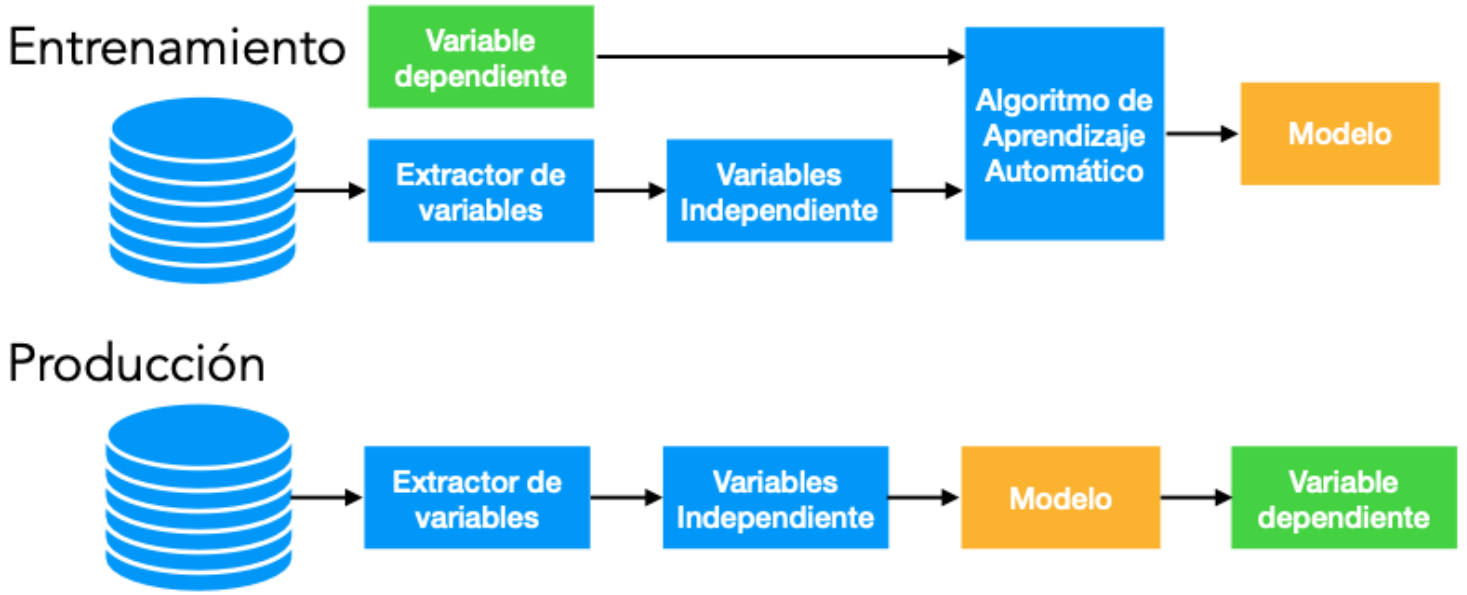
¿Validación sobre todo el dataset?

TIPOS DE MODELOS



Rendimiento de los modelos

MODELOS EN PRODUCCIÓN



Problemas:
- Model decay/Model drift (envejecimiento del modelo): peores rendimientos con el tiempo (cambios significativos en el entorno del entrenamiento inicial) → Ojo en sistemas dinámicos.
- Data/Feature/Concept drift: deriva de datos, variables o relaciones entre variables → El modelo envejece porque las distribuciones cambian con el tiempo, las variables ya no representan los mismo o sus interacciones no son las mismas.

Mitigación → Ingreso de nuevos datos, reentrenamiento y monitorización continua de los rendimientos.

UAX Universidad Alfonso X el Sabio

GRACIAS

DR. ARIEL CARIAGA-MARTINEZ

CIENCIA DE DATOS

ACARIMAR@UAX.ES

UAX Universidad Alfonso X el Sabio

TRABAJOS

GRADO EN BIOMEDICINA

DR. ARIEL CARIAGA-MARTÍNEZ

UN “NUEVO” MODELO

- Formar equipos (4 personas).
- Utiliza los datasets indicados y sus diccionarios de datos.
- BUSCA INFORMACIÓN SOBRE CÓMO GENERAR MODELOS BÁSICOS DE CLASIFICACIÓN Y REGRESIÓN.
- Seguramente ya conoces como hacerlo: busca una modelización retadora y explica qué significa cada *outcome*.
- APLICA LOS PRINCIPIOS BÁSICOS QUE VIMOS EN LA TEORÍA: POR AHORA NO HACE FALTA QUE HAGAS UN “MACHINE LEARNING PERFECTO”. SOLO VAMOS A ACERCARNOS AL MODELO.
- Genera visualizaciones: ¿qué ”historia” podrías contar tras el modelado?