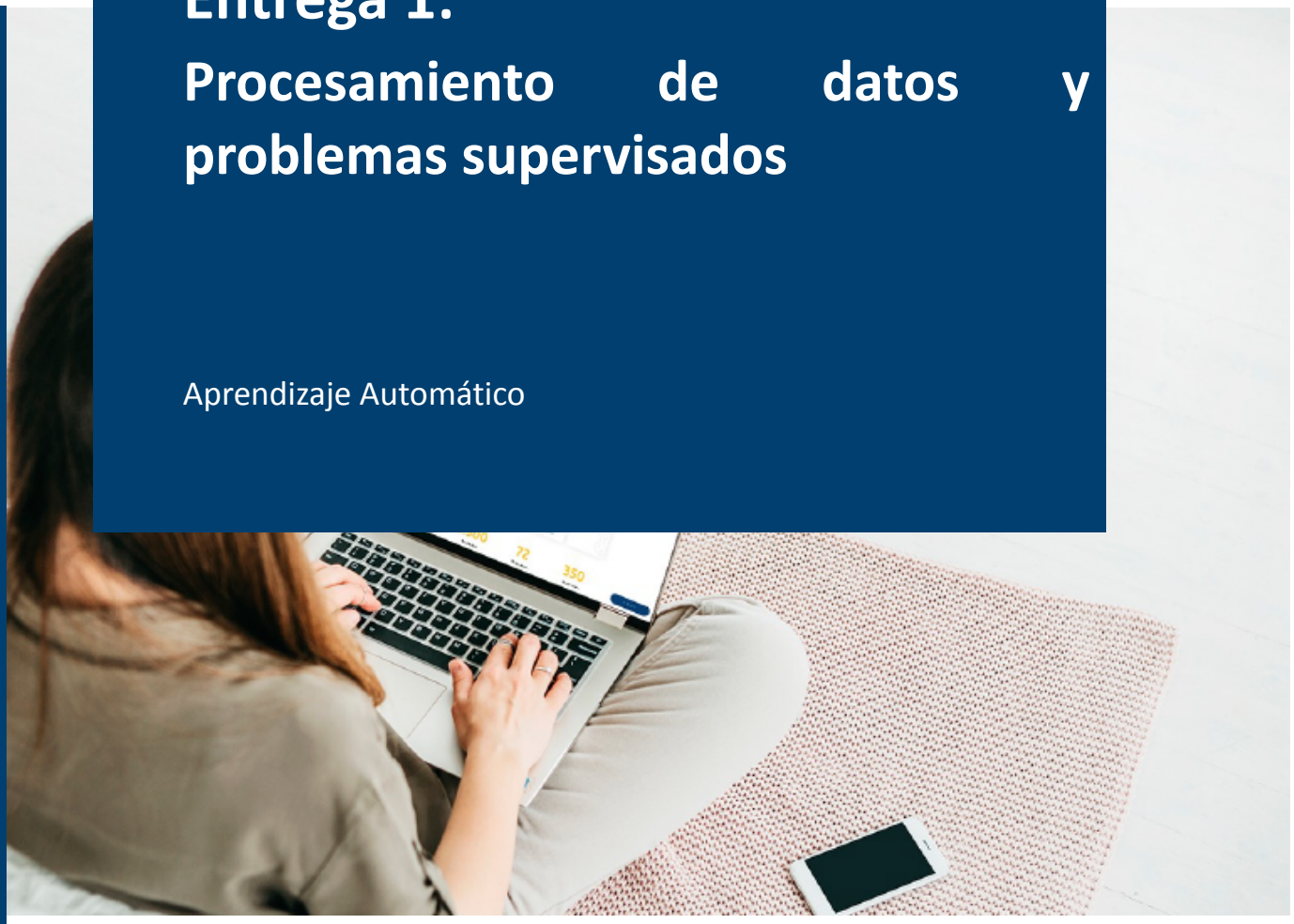
OPENUAX
UNIVERSIDAD ALFONSO X EL SABIO

# Entrega 1:
# Procesamiento de datos y problemas supervisados

Aprendizaje Automático



# Objetivo

El objetivo de esta entrega es poner en práctica los conocimientos adquiridos en las Unidades 1 y 2. Se van a poner a tu disposición dos conjuntos de datos que contienen información sobre dos problemas, los cuales se pretenden resolver con aprendizaje automático. En el primero, se busca estimar el precio de venta de un coche de segunda mano, mientras que en el segundo se quiere predecir el resultado de una campaña de marketing telefónico. Ambos conjuntos de datos requieren en primer lugar un análisis y procesamiento de la información, para posteriormente entrenar algoritmos de aprendizaje supervisado. En esta entrega tendrás que implementar la solución a ambos problemas utilizando Python.

# 1. Ejercicio 1: estimación del precio de coches usados

En el primer ejercicio, queremos **estimar el precio de venta de un coche de segunda mano** en base a sus características. Disponer de un algoritmo predictivo capaz de resolver esta tarea puede tener multitud de aplicaciones: sugerir precios en portales de compra-venta online, tasación automática, etc.

Para resolver este problema se ha recopilado un conjunto de datos con ventas históricas de coches usados, incluyendo el precio real al que se vendió. El dataset contiene las siguientes 11 columnas:

- **ID:** identificador del coche
- **marca**: marca del vehículo: Audi, BMW, Skoda…
- **modelo**: el modelo del coche. Por ejemplo, A7
- **fecha**: es la fecha original de fabricación del vehículo, e informa sobre su antigüedad
- **tipo_cambio**: indica el tipo de cambio que tiene el coche: manual, automático…
- **total_km**: es el kilometraje actual del coche
- **tipo_combustible:** indica el tipo de combustible que usa el coche: diesel, gasolina (*petrol*)...
- **consumo**: es el consumo de combustible del coche. Se mide en litros por cada 100 km
- **tipo_motor**: refleja el tamaño del motor, y está relacionado con la potencia del mismo. Por ejemplo, 3.0
- **tasa**: es el impuesto que es necesario pagar para poder circular con el coche. Se mide en euros (€)
- **precio**: es el precio en euros al que se vendió el vehículo de segunda mano. **Esta es la variable objetivo a predecir.**

En el espacio correspondiente a la Entrega 1 en el Campus Virtual encontrarás 2 ficheros que contienen los datos para este problema:

- El fichero ***dataset_coches_train.csv*** contiene el conjunto de entrenamiento, es decir, el conjunto de datos para ajustar y optimizar los algoritmos de aprendizaje supervisado. Contiene en total 4,960 registros.
- El fichero ***dataset_coches_test.csv*** contiene el conjunto de validación, es decir, el conjunto de datos para validar y comprobar la capacidad de generalización del modelo entrenado. Contiene un total de 2,672 registros.

En este primer problema se pide **implementar y entrenar un algoritmo de aprendizaje supervisado** que sea capaz de predecir la variable **precio**. Se considerará que el modelo es satisfactorio si se obtiene un **R2 superior a 0.9** en el conjunto de validación.

Adicionalmente, y con el fin de entender mejor y profundizar en los datos disponibles, se pide resolver las siguientes cuestiones:

1. Representar gráficamente la distribución de la variable precio en el conjunto de entrenamiento
2. ¿Cuál es la marca más cara en promedio? ¿Y la más barata?
3. Representar gráficamente la dependencia entre el precio y el kilometraje
4. Calcular las variables más significativas, ya sea mediante un test estadístico o analizando el modelo entrenado
5. Un amigo quiere vender un audi A7 de 2020 con 5000 km, cambio automático, combustible híbrido, consumo de 5.5 l/100km y motor 4.0. La tasa de circulación es de 200€. ¿A cuánto debería venderlo?

# 2. Ejercicio 2: predicción del resultado de una campaña de marketing telefónica

Un banco realiza una campaña de marketing telefónica para ofrecer un depósito a plazo fijo a sus clientes. La campaña se ha llevado a cabo durante 3 años y se han recopilado los datos de la misma. Ahora que se acerca el cuarto año, el banco quiere mejorar los resultados de la campaña y nos pide que desarrollemos una solución para predecir si un cliente contratará su producto o no. De esta forma, se pueden gestionar mejor las llamadas al centrarse en los usuarios con mayor propensión.

El banco ha aglutinado la información en un dataset que contiene variables sobre el cliente, su situación financiera, los datos de la campaña o indicadores macroeconómicos. En total se dispone de las siguientes 17 columnas:

- **edad:** es la edad del cliente, en años
- **empleo**: es el tipo de empleo del cliente
- **estado**: es el estado civil del cliente: soltero, casado, divorciado, etc.
- **educación:** es el nivel de estudios del cliente
- **impago:** indica si el cliente tiene algún impago pendiente
- **hipoteca:** indica si el cliente tiene contratada una hipoteca

- **préstamo**: indica si el cliente tiene contratado un préstamo
- **tipo_contacto:** indica si se contactó con el cliente por teléfono fijo o móvil (*cellular*).
- **mes:** el mes en el que se contactó con el cliente
- **dia_semana:** el día de la semana en el que se contactó con el cliente
- **contactos_actual**: indica cuántas veces se ha contactado con el cliente en la campaña actual
- **contactos_anterior:** indica cuántas veces se ha contactado con el cliente en la campaña anterior
- **resultado_anterior:** refleja el resultado de la campaña anterior: *success* si contrató el producto, *failure* si lo rechazó o *nonexistent* si no se contactó con el cliente
- **tasa_var_empleo_3m**: indicador macroeconómico que indica la variación en la tasa de empleo trimestral
- **euribor_3m**: indicador macroeconómico que indica el valor del Euribor trimestral
- **ipc_1m**: indicador macroeconómico que indica el valor del IPC (Índice de Precios al Consumidor) mensual
- **target**: es el resultado de la campaña: *yes* si contrató el préstamo y *no* en caso contrario. **Esta es la variable objetivo a predecir.**

En el espacio correspondiente a la Entrega 1 en el Campus Virtual encontrarás 2 ficheros que contienen los datos para este problema:

- El fichero ***dataset_marketing_train.csv*** contiene el conjunto de entrenamiento, es decir, el conjunto de datos para ajustar y optimizar los algoritmos de aprendizaje automático. Contiene 32,652 registros correspondientes a los dos primeros años de campaña.
- El fichero ***dataset_marketing_test.csv*** contiene el conjunto de validación, es decir, el conjunto de datos para validar y comprobar la capacidad de generalización del modelo entrenado. Contiene 8,536 registros correspondientes al tercer año de campaña.

Para el segundo problema se pide **implementar y entrenar un algoritmo de aprendizaje supervisado** que sea capaz de predecir la variable **target**. El banco considerará que el modelo es satisfactorio si se obtiene una **tasa de acierto (accuracy) superior a 0.9** en el conjunto de validación.

Adicionalmente, y para poder entender mejor los datos disponibles, se pide resolver las siguientes cuestiones:

1. Calcular el ratio de conversión (porcentaje de clientes que contratan) de la campaña en el conjunto de entrenamiento
2. ¿Cómo influye el día de la semana de contacto en el resultado de la campaña? ¿Y el mes?
3. Transformar, al menos, las variables estado y resultado_anterior utilizando one-hot encoding
4. Representar gráficamente la curva ROC del modelo y calcular su AUC en el conjunto de validación

5. El banco quiere optimizar los costes de la campaña utilizando el modelo que acabas de entrenar. Para ello, te indica que cada llamada que se realiza tiene un coste de 1€ para ellos, y que el beneficio obtenido al contratar el producto es de 100€. Por tanto, el balance neto (beneficios - gastos) de una campaña es:

   Balance = 100€ x Clientes que contratan - 1€ x Clientes contactados

   Para maximizar el balance se necesita que el mayor número posible de llamadas terminen en contratación. Es decir, que la precisión sea lo más alta posible. Por ejemplo: si se llama a 10 clientes y ninguno contrata el balance es -10€, y si contratan 5 el balance sería 490€.

   El banco va a utilizar tu modelo para decidir a qué clientes llama y a cuáles no: llamará sólo a aquellos que tengan una probabilidad de contratación superior a un cierto umbral. Utilizando el conjunto de validación, deberás hallar cuál es el umbral que maximiza el balance de la campaña teniendo en cuenta los beneficios y los gastos.

   Nota: El conjunto de validación contiene 8,536 clientes contactados, pero ahora lo que vas a hacer es simular qué hubiera pasado si se utilizase tu modelo. Es decir, para resolver esta pregunta debes calcular tanto el número de clientes contactados, en función de las probabilidades de tu modelo, como los que contratan.

## 3. Instrucciones de entrega

- **Formato**: Se debe entregar en formato cuaderno de Jupyter (fichero con extensión .ipynb). Se puede entregar un único cuaderno con ambos ejercicios o dos cuadernos distintos.
- **Ejecución**: El cuaderno con la solución debe poder ejecutarse de principio a fin, cargando los datasets desde ficheros locales.
- **Entrega**: Subir los ficheros al buzón de entrega correspondiente a la Entrega 1 en el Campus Virtual.
- **Intentos**: Se pueden hacer intentos ilimitados hasta la fecha de entrega. Solo se evaluará la última versión subida.
- **Fecha de entrega**: El plazo de entrega finaliza el lunes 19 de diciembre de 2022 a las 16:00 (hora española).

# 4. Evaluación

La calificación final de la entrega será la suma de las calificaciones individuales de cada ejercicio, los cuales se evaluarán entre 0 y 5.

Cada ejercicio se evaluará de acuerdo a los siguientes criterios:

- (2 puntos) Conseguir entrenar un algoritmo que supera las métricas establecidas en el conjunto de validación (R2 > 0.9 en el ejercicio 1, tasa de acierto > 0.9 en el ejercicio 2).
- (1.5 puntos) Contestar correctamente a las 5 preguntas planteadas. Las preguntas deben contestarse explícitamente con Python, no se valorarán otro tipo de respuestas.
- (1 punto) Analizar y experimentar en profundidad. Se valorará positivamente que se hayan llevado a cabo diferentes análisis y experimentos/entrenamientos, independientemente del resultado. Por ejemplo: visualizaciones, creación de nuevas variables, entrenar varios modelos, uso de normalización o selección de variables, evaluar con diferentes métricas, etc.
- (0.5 puntos) Presentación y claridad de la entrega, tanto en el código como en la explicación

La Entrega 1 supondrá un 20% de la nota final de la asignatura.

No se evaluarán las entregas que se suban después de la fecha límite establecida en el aula virtual.


OPENUAX
UNIVERSIDAD ALFONSO X EL SABIO
WELCOME
TO
UAX
GRACIAS
UAX.COM