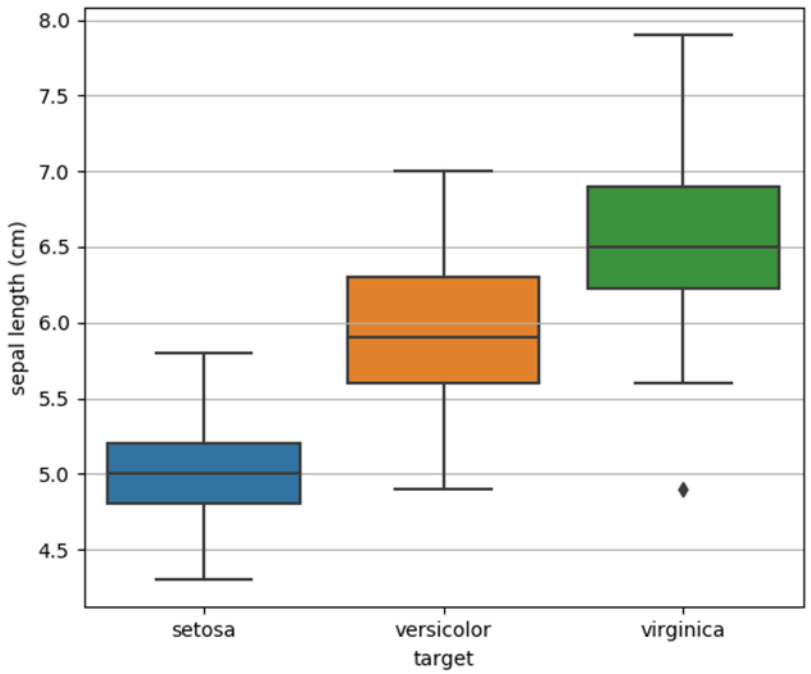
Actividad. Visualización y análisis exploratorio

Unidad	4
Entrega	Documento en Jupyter Netbook

1. Enunciado

La actividad está organizada en ejercicios y se evalúa sobre un máximo de 10 puntos. Todos los ejercicios tienen el mismo peso en la evaluación de la actividad. Un ejercicio puede contener varias preguntas o apartados, en cuyo caso la puntuación del ejercicio se repartirá de forma equitativa entre las preguntas o apartados que lo componen.

1. El siguiente boxplot representa los conjuntos de valores de la longitud del sépalo de la flor de tres tipos distintos de iris. Responded verdadero o falso a las siguientes afirmaciones. Basad vuestras respuestas en la información que veis en la gráfica, sin usar el dataset original ni calcular los valores con código.



   a. Las medianas de la longitud del sépalo para las variedades *Versicolor* y *Virginica* son iguales.

   b. Todos los elementos de la variedad *Setosa* tienen una longitud del sépalo inferior a la de cualquier elemento de la variedad *Virginica*.

   c. El mínimo valor de longitud del sépalo de la variedad *Virginica* es más de medio centímetro más largo que el mínimo valor de la longitud del sépalo para la variedad *Versicolor*.

d. Podemos asegurar que el 75% de los elementos de la variedad *Virginica* tienen una longitud del sépalo superior a la de cualquier elemento de la variedad *Setosa*.

e. Solo el 40% de los elementos de la variedad *Setosa* tienen una longitud del sépalo mayor que el mínimo de la longitud del sépalo de la variedad *Versicolor*.

f. Solo la variedad *Virginica* presenta un valor anómalo de la longitud del sépalo.

2. La siguiente tabla muestra los valores estadísticos de cuatro variables que consideramos para la clasificación de variedades de iris en función de las características de sus flores. Responde a las siguientes preguntas analizando los valores de la tabla:

	sepal length (cm)	*sepal width* (cm)	*petal length* (cm)	*petal width* (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

a. ¿De cuántos elementos se compone el dataset?

b. ¿Cuál es el rango intercuartílico de la longitud del pétalo?

c. Responde "Verdadero" o "Falso" justificando la respuesta: la media es siempre igual a la mediana.

d. Responde "Verdadero" o "Falso" justificando la respuesta: la media es siempre superior a la mediana.

e. ¿Cuál es la variable con menor desviación típica? ¿Cuál tiene menor desviación típica comparado con su media?

f. Responde "Verdadero" o "Falso" justificando la respuesta: como medida de dispersión de datos, a mayor rango de una variable, corresponde siempre mayor desviación típica.

3. Utiliza el siguiente enlace para descargar el California *Housing Prices* Dataset. Utiliza un Jupyter Notebook y el paquete Pandas para abrir el archivo y presentarlo en un DataFrame llamado df_house, donde el nombre de las columnas debe corresponder con el nombre real de las variables.

<https://www.kaggle.com/camnugent/california-housing-prices>

Elimina la variable "ocean_proximity" de df_house. Elimina las instancias que contengan valores faltantes. Crea un array de nombre y_house con los valores de la variable "median_house_value". Elimina la variable "median_house_value" del DataFrame df_house.

a. Utiliza el método describe() de Pandas para obtener un resumen de las estadísticas de las variables del DataFrame df_house. En base a los valores de mediana, percentil 25 y percentil 75, ¿qué variables crees que siguen una distribución distinta a la distribución normal?

b. Dibuja los histogramas de cada una de las variables del dataset. ¿Coincide la predicción que has hecho basada en los percentiles con los resultados gráficos que has obtenido con los histogramas? Algunos algoritmos de Machine Learning funcionan mejor cuando sus variables predictivas siguen una distribución normal. ¿Qué tipos de algoritmos crees que tienen este requerimiento? ¿Qué tipo de transformaciones puedes aplicar a las variables de este dataset para que las variables nuevas sigan, aproximadamente, una distribución normal?

c. Dibuja el qq-plot para cada una de las variables del dataset, donde cada variable ha sido normalizada de forma que tiene media cero y varianza uno. ¿Cómo podemos distinguir las variables que siguen distribuciones normales de las que no las siguen utilizando este tipo de gráficos? ¿Qué variables no siguen una distribución normal de acuerdo a los gráficos qq-plot que has dibujado?

d. Crea un DataFrame nuevo, llamado df_trans, donde las variables originales han sido transformadas de forma que cada una de ellas sigue, aproximadamente, una distribución normal. Para ello utiliza, cuando sea posible, una transformación de Box-Cox. De acuerdo al parámetro de la transformación para cada variable ¿cuál sería, aproximadamente, la transformación que sufren cada una de ellas (logaritmo, raíz cuadrada, …)? Nombra las variables de df_trans como las variables del df original, añadiendo el sufijo "_trans".

e. Crea un nuevo array y trans a partir del array y_house, de forma que los elementos de y_trans sigan, aproximadamente, una distribución normal. Utiliza una transformación de Box-Cox. Representa, para cada variable transformada, el qq-plot correspondiente. Recuerda que debemos normalizar la variable de forma que tenga media cero y varianza uno. ¿Siguen las variables transformadas una distribución normal?

4. Utiliza los DataFrames df_house y df_trans y los arrays y_house e y_trans, obtenidos en el ejercicio anterior para realizar los siguientes apartados:

a. Sobre el array y_house, utiliza el método de z-score para señalar las instancias que muestren valores atípicos. Sobre el array y_trans, utiliza el método z-score para señalar las instancias que muestren valores atípicos. ¿Coinciden las instancias encontradas en ambos arrays? ¿Cuál crees que es el motivo?

b. Realiza un análisis representando gráficamente un gráfico de caja para los arrays y_house e y_trans. ¿Muestran la misma distribución de valores atípicos? ¿Cuál crees que es el motivo? Utilizando la definición del gráfico de cajas obtén las instancias de y_house e y_trans que muestren valores atípicos. ¿Son las mismas que las obtenidas en el apartado anterior?

c. Utiliza la clase LocalOutlierFactor de Scikit Learn para declarar un objeto llamado lof. Entrena este objeto sobre df_house para encontrar las instancias que muestren valores atípicos. ¿Cuáles son estas instancias?

d. Entrena el objeto lof sobre df_trans para encontrar las instancias que muestren valores atípicos. ¿Son las mismas instancias que en el apartado anterior?

e. Utiliza la clase IsolationForest para crear un objeto llamado isf. Entrena este objeto sobre df_house para encontrar las instancias que muestren valores atípicos. ¿Son las mismas instancias que en el apartado c? Repite el proceso entrenando sobre df_trans, ¿obtienes las mismas instancias que las que obtuviste en el apartado d?

2. Detalles de la entrega

- Las respuestas de la actividad se deberán entregar en un Jupyter Notebook en el que se haya respondido a cada apartado en una celda independiente, en el orden de las preguntas de este documento. El código de cada celda se debe poder ejecutar para comprobar las respuestas aportadas. Las preguntas teóricas se deben responder en una celda de formato texto.
- Subir de forma grupal el documento a la actividad en el campus virtual.