

## PR2.3. ÁRBOLES DE DECISIÓN

### Descripción de la tarea

---

Realizar estudios de aprendizaje supervisado utilizando árboles de decisión, tanto para clasificación como para regresión. Para ello en esta tarea se usarán 4 datasets (los tres de la práctica anterior y uno que introduce series temporales).

1. <https://www.kaggle.com/yasserh/wine-quality-dataset> . Clasificación de la calidad del vino según alguno de sus parámetros (ph, densidad, sulfatos, alcohol, azúcar, etc.). El dataset contiene una serie de datos sobre un vino y su calidad (de 0 a 10). Este problema es de clasificación puesto que consiste en clasificar el vino en concreto en su clase de calidad (valores discretos del cero al 10)
2. <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>. Clasificación binaria. El objetivo es clasificar entre presencia de enfermedad del corazón o no a partir de una serie de atributos (edad, sexo, colesterol, etc.). Para ello el dataset contiene una serie de datos de pacientes etiquetados si tienen enfermedad o no.
3. <https://www.kaggle.com/schirmerchad/bostonhousingmlnd>. Predicción de precios de las casas de Boston. El dataset contiene una serie de datos de casas. El objetivo es, dada una serie de parámetros, predecir el precio que debe tener la casa (MEDV).
4. <https://www.kaggle.com/marklvl/bike-sharing-dataset>. Predicción de número de bicicletas alquiladas por cada hora. El dataset contiene una serie de datos como temperatura, humedad, si es día laborable, etc. El objetivo es predecir cuántas bicicletas se van a alquilar dependiendo de esa serie de parámetros.

Para cada uno de ellos realizar un documento de google colab con los siguientes epígrafes y tareas:

- Importación de librerías necesarias
- Preproceso
  - Importación de los datos del dataset
  - Mostrar las primeras y últimas filas del dataframe importado
  - Mostrar parámetros estadísticos de los datos (media, desviación típica, cuartiles, etc.)
  - Mostrar un mapa de calor que indique la correlación entre variables
  - Seleccionar las características a tener en cuenta en el estudio
  - Separar datos entre datos de entrada y etiquetas (resultados)

- Separar datos entre entrenamiento y prueba (usando un 80% para entrenamiento y 20% para test)
- Entrenamiento y predicción
  - Elegir, instanciar (eligiendo unos valores concretos, por ejemplo profundidad del árbol 3) y entrenar el modelo
  - Realizar una predicción con los datos de prueba
  - Mostrar el árbol de decisión resultante
  - Mostrar la importancia de cada atributo en el árbol resultante
  - Intenta guardar el modelo de predicción ya entrenado usando dump ([https://scikit-learn.org/stable/modules/model\\_persistence.html](https://scikit-learn.org/stable/modules/model_persistence.html))
- Evaluación
  - Para los problemas de clasificación
    - Mostrar el porcentaje de elementos correctamente clasificados
    - Mostrar la predicción realizada (imprimir la variable con la predicción)
    - Representar gráficamente la clasificación obtenida (matriz de confusión)
  - Para los problemas de regresión
    - Mostrar el error cuadrático medio (mean\_squared\_error)
    - Mostrar el error absoluto medio (mean\_absolute\_error)
    - Representar gráficamente los valores predichos con los valores reales.

```
1 # x axis for plotting
2 import numpy as np
3 import matplotlib.pyplot as plt
4 xx = np.stack([i for i in range(y_test.shape[0])])
5 plt.plot(xx, y_test, c='r', label='data')
6 plt.plot(xx, y_pred, c='g', label='prediction')
7 plt.axis('tight')
8 plt.legend()
9 plt.show()
```

- Optimización de hiperparámetros
  - Calcula la combinación de parámetros óptima (profundidad de árbol y criterio). Para ello realiza ejecuciones con cada uno de los valores del criterio para los valores de profundidad de árbol de 1 a 15.
  - Cada ejecución anterior se deberá hacer usando validación cruzada (por ejemplo n\_splits =

5). Con ello obtendremos una medida de bondad del modelo (accuracy\_score o mean\_absolute\_error)), como lo ejecutaremos 5 veces, calcularemos la media de esas 5 ejecuciones.

- Finalmente los parámetros elegidos serán los que den mejor media de esas medidas anteriormente nombradas.
- Una vez obtenidos esos parámetros óptimos los aplicaremos al problema en cuestión y mostraremos los resultados.

**Documentación a entregar.** Se entregará en moodle centros un archivo (cuaderno de júpiter o google colab) para cada uno de los ejercicios con el desarrollo de las tareas anteriormente mencionadas. Además se entregará un fichero pdf con unas conclusiones (para cada ejercicio) que describan todas las fases del proceso de aprendizaje supervisado que se han ido desarrollando en esta tarea (explicación de qué se ha realizado)

## Evaluación

---

Los criterios de evaluación de la tarea son a, b, c, d, e, f, g, h del RA3. Aplica algoritmos de aprendizaje supervisado, optimizando el resultado del modelo y minimizando los riesgos asociados.

Para evaluar la práctica se puntúan los siguientes apartados:

- Clasificación de vinos (2,5 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.
- Clasificación presencia de enfermedad de corazón (2,5 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.
- Predicción del precio de las casas en Boston (2,5 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.
- Predicción del número de bicicletas alquiladas (2,5 puntos)
  - Preproceso; Entrenamiento y predicción; Evaluación; Optimización; Conclusiones.

## Rúbrica por cada problema

	100%	75%	50%	25%	0%
<b>Preproceso</b>	Proporciona los datos etiquetados al modelo, importándolos y realizando todas la modificaciones necesarias.	Proporciona los datos etiquetados al modelo, hay algunos errores menores en la importación, pero los datos se adaptan	Proporciona los datos etiquetados al modelo, hay algunos errores en la importación o no se realizan modificaciones	Proporciona los datos etiquetados al modelo, hay errores en la importación y no se realizan modificaciones	No realiza la importación de datos, o la realiza de forma errónea
<b>Entrenamiento y predicción</b>	Realiza el entrenamiento del modelo y la predicción con los datos de test	Realiza el entrenamiento del modelo y la predicción con los datos de test. Pero existen algunos errores menores	Realiza el entrenamiento correctamente pero no la predicción	Realiza el entrenamiento pero no la predicción y existen algunos errores en el entrenamiento	No realiza el entrenamiento del modelo, ni la predicción o presentan errores
<b>Evaluación</b>	Se realiza una evaluación del modelo con los datos. Además se representa de diversas formas, gráficamente en texto	Se realiza una evaluación del modelo con los datos. Sólo se presenta un gráfico	Se realiza una evaluación del modelo con los datos. Se muestra un gráfico pero con algunos errores menores	Se realiza una evaluación del modelo con los datos. Sólomente en texto	No se realiza la evaluación o presenta errores importantes
<b>Optimización</b>	Se realiza una optimización de hiperparámetros o una comparativa entre distintos modelos de la misma familia.	Se realiza una optimización de hiperparámetros o una comparativa pero existen algunos errores menores o no está suficientemente detallado	Se realiza una optimización de hiperparámetros o Hay algún error en los los valores óptimos o en la comparativa.	Se realiza una optimización de hiperparámetros o una comparativa. Existen errores importantes, no se consiguen los valores óptimos o la comparativa es escasa o irrelevante	No se realiza la optimización ni se prueban diversos modelos de la misma familia
<b>Conclusiones</b>	Realiza unas conclusiones que describen todas las fases del proceso de solucionar un problema con aprendizaje supervisado	Realiza unas conclusiones que describen la mayoría de las fases del proceso de solucionar un problema con aprendizaje	Realiza unas conclusiones que describen al menos la optimización de parámetros y la evaluación de los resultados obtenidos	Realiza unas conclusiones que describen alguna de las fases del proceso de solucionar un problema con aprendizaje	No realiza conclusiones o son erróneas o irrelevantes

## Refuerzo y ampliación

---

### Refuerzo.

1. Realiza el estudio de aprendizaje supervisado (usando árboles de decisión) para el problema de clasificación de los distintos tipos de flores iris según el ancho y el alto del pétalo y sépalo. Para ello utiliza el dataset iris (<https://archive-beta.ics.uci.edu/dataset/53/iris>)

### Ampliación.

- Realiza un estudio de aprendizaje supervisado con el siguiente dataset (para predecir el precio de ordenadores) <https://www.kaggle.com/muhammetvarl/laptop-price>
- Consulta el dataset <https://www.kaggle.com/vivianyfwang/capital-bike-share-2018-20218> e indica que preprocesamiento crees que se debería hacer. Realiza ese preprocesamiento, entrenamiento y predicción y compara dichas predicciones con las obtenidas en el ejercicio4.
- Realiza un estudio de aprendizaje supervisado con el siguiente dataset (para predecir el salario de un jugador de NBA dadas sus estadísticas) <https://www.kaggle.com/aishjun/nba-salaries-prediction-in-20172018-season>
- Consulta los siguientes enlaces que contienen datasets para machine learning <https://archive-beta.ics.uci.edu/datasets> y <https://www.kaggle.com/datasets>.
  - Elige algún/algunos datasets y plantea y resuelve un problema de clasificación y otro de regresión usando árboles de decisión.
- Consulta con detenimiento el siguiente enlace <http://www.cs.us.es/~fsancho/?e=104>
  - Trata de implementar el algoritmo ID3 siguiendo el pseudocódigo mostrado.
    - Un ejemplo de implementación se puede consultar en <https://programmerclick.com/article/85151567971/>