Máster Universitario en
Lógica, Computación e Inteligencia Artificial
Aprendizaje Automático

Apellidos:.....

Nombre :.....

En este cuestionario se espera que el alumno profundice en algunos aspectos de la implementación de los árboles de decisión que no se han visto en clase. Se espera que el alumno presente **el mejor análisis posible** dependiendo de su formación básica y de su experiencia con sistemas de aprendizaje automático.

Para ello, cada alumno debe elegir una base de datos del repositorio de la Universidad de California (accesible en https://archive.ics.uci.edu/ml/datasets.php) y realizar sobre él el siguiente estudio.

1. Reproducir los pasos que se han visto en clase sobre el conjunto iris, en particular:

   a) Carga del fichero en python y probablemente, eliminación de las filas que no tengan todos los datos. Esto último no se ha visto en clase. Se sugiere mirar la documentación de la librería *pandas*.

   b) Realiza un pequeño examen exploratorio creando la *scatter matrix* y el mapa de calor *(heatmap)* de los coeficientes de correlación de Pearson entre las variables.

   c) Divide la base de datos en conjunto de entrenamiento y prueba y crea el correspondiente árbol de decisión sobre le conjunto de entrenamiento.

   d) Representa gráficamente el árbol obtenido.

   e) Encuentra la media de rendimiento *(score)* del árbol sobre los conjuntos de entrenamiento, prueba y total.

2. La parte realmente interesante de este cuestionario empieza ahora, donde se espera que el alumno demuestre su **autonomía** y **capacidad de profundización** en la materia a partir de los conceptos básicos. Se espera que el alumno continúe con el análisis del conjunto de datos y presente el estudio **más completo posible**. Algunas posibles líneas para profundizar son las siguientes (queda a elección del alumno elegir una, elegir variar, elegirlas todas o ampliarlas con otras posibles líneas no contempladas en este cuestionario):

   a) **Prepoda. Mínimo número de muestras en nodos internos** Podemos limitar el número de muestras que debe tener un nodo para considerarlo un nodo interno del árbol. Este valor se indica con el parámetro *min_samples_split*. Si un nodo no tiene suficientes muestras asociadas entonces será considerado un nodo hoja cuyo valor de clasificación será el mayoritario entre sus muestras. ¿Mejora con este método el rendimiento de nuestro árbol?

   b) **Prepoda. Mínimo grado de impureza** Este criterio detiene el desarrollo del árbol cuando la proporción de la clase dominante en el conjunto

de muestras asociado es muy alta en comparación con las de las otras clases. Este valor se indica con el parámetro *min_impurity_split* como un valor comprendido entre 0 y 1. Si la impureza de un nodo no alcanza el mínimo, entonces será considerado un nodo hoja cuyo valor de clasificación será el mayoritario entre sus muestras. ¿Mejora con este método el rendimiento de nuestro árbol?

c) **Validación cruzada** ¿Mejora el rendimiento de nuestro árbol si usamos validación cruzada? ¿Y si usamos validación cruzada con estratificación?

d) ¿Qué otros métodos podríamos utilizar para obtener el *mejor* árbol de decisión posible?

Nota: Los métodos *ensemble* basados en árboles de decisión *(Random forest, Gradient boosted, etc)* se estudiarán en la asignstura *Inteligencia Artificial para la Ciencias de los Datos* del segundo cuatrimestre.