Clase N°3. Árboles de clasificación

Los **árboles de clasificación** son aquellos donde la variable predictoria es de tipo categórica siendo sus factores variables numéricas. Para construir su estructura se implementa el mismo método de *división recursiva binaria* utilizado en los árboles de regresión, con la diferencia que en los nodos hojas se recogen variables categóricas y las métricas utilizadas analizan el grado de exactitud en la predicción. Entre los métodos más utilizados se encuentran el **índice Gini** y la **entropía**.

Índice Gini

El indicador mide el nivel de pureza en cada nodo al momento de realizar la partición binaria. Los nodos con mayor pureza expresan mejor resultado en la predicción, mientras que los **nodos impuros** reflejan resultados con menor exactitud en sus pronósticos. Por ejemplo un nodo impuro está formado por elementos de distintas clases y un **nodo puro** está constituido por objetos de un mismo tipo.

El valor de índice varía entre 0 y 1. Los resultados próximos a 1 indican mayor grado de impureza en el nodo, mientras que los valores próximos a 0 representan mayor nivel de pureza. Al recorrer el árbol de clasificación, los nodos van disminuyendo su nivel de impureza en cada partición binaria hasta llegar a los nodos terminales donde el nivel de pureza es óptimo.

Entropía

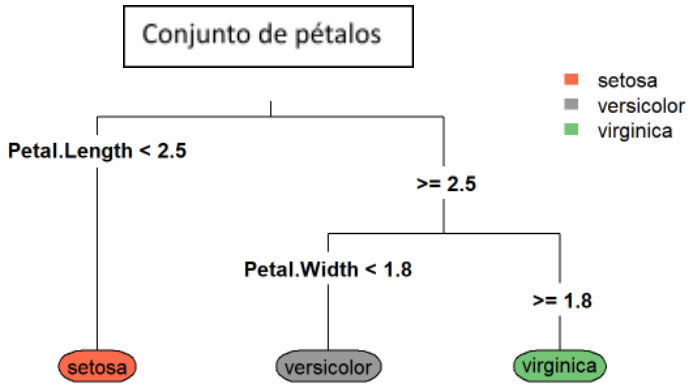
La entropía es un concepto muy utilizado en física y en particular en termodinámica para medir sistemas desordenados o en desequilibrio. En los árboles de clasificación, la entropía mide el **grado de desorden** en cada nodo al momento de realizar una partición, es decir, si un nodo presenta un alto nivel de impureza también representa un alto grado de entropía. Por otra parte, los nodos puros, los cuales están formados por elementos de una misma clase presentan un **alto grado de orden**.

Al igual que en índice Gini, la entropía varía entre 0 y 1. Los resultados próximos a cero reflejan un alto grado de orden, en tanto que los valores próximos a uno indican un gran nivel de desorden.

De la misma forma que en el árbol de regresión, se cumplen las mismas etapas de un modelo predictivo: preprocesamiento, partición, entrenamiento, evaluación y validación.

Ejemplo: clasificación de un conjunto de flores

Clasificar un conjunto de flores como: setosa, versicolor o virginica en relación al largo y ancho del pétalo.



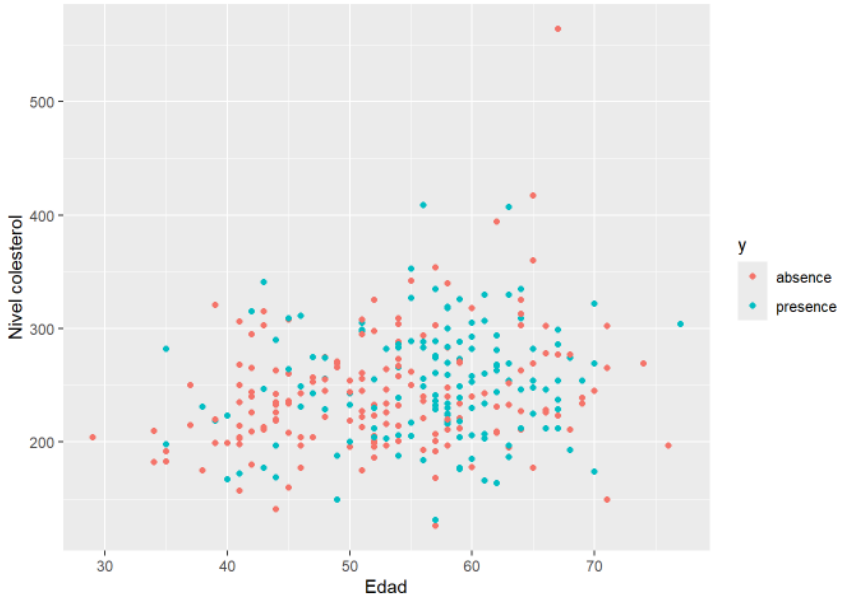
Ejemplo 2: Predecir si un paciente es propenso a sufrir una enfermedad cardíaca.

El objetivo es crear un árbol de clasificación para predecir la variable Y (`target`) definida como:

$$Y = \begin{cases} 1 & \text{si paciente SI sufre una enfermedad cardíaca} \\ 0 & \text{si paciente NO sufre una enfermedad cardíaca} \end{cases}$$

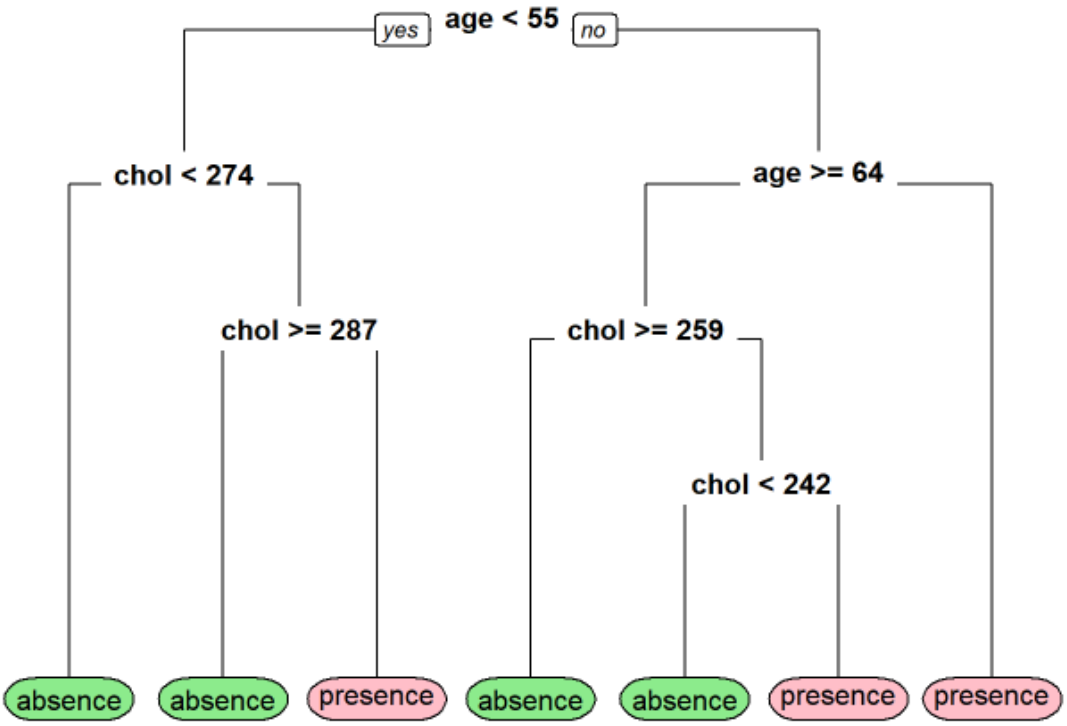
En función de las variables observables: edad "`age`" y nivel de colesterol "`chol`" mg/dl (unit). Arbitrariamente se decide usar los nombres "`absence`" y "`presence`" en lugar del 0 y 1 para facilitar su interpretación.

A partir de la base de datos, se puede graficar la relación entre la edad, el nivel de colesterol y la ausencia o presencia de la enfermedad cardíaca, como muestra en el sig. gráfico:



En el gráfico se puede notar que al aumentar la edad y al incrementar el nivel de colesterol, existe mayor probabilidad de presencia de enfermedades cardíacas. Luego al volcar esta información en un árbol de clasificación, resulta:

Árbol de clasificación



Para obtener la tasa de clasificación correcta podemos usar una **matriz de confusión**, la cual evalúa los aciertos en la predicción en la diagonal principal y los fallos en la predicción en los demás valores. En este caso la matriz de confusión es:

	Clasificacion	
Verdadero	absence	presence
absence	124	40
presence	56	83

Donde la taza de aciertos es del 68% al sumar los valores de la diagonal principal divididos por el total de datos.

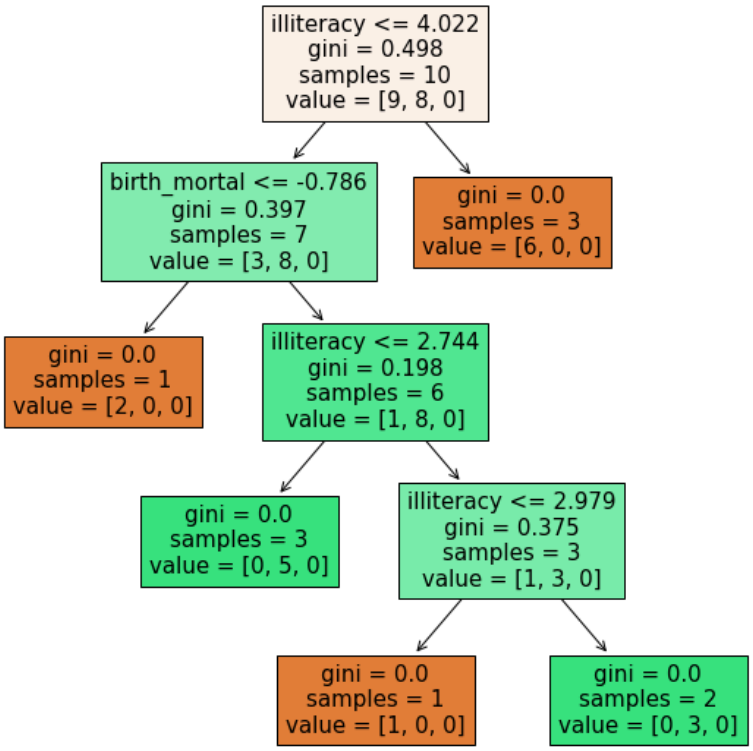
Predicción:

Supongamos que tenemos un paciente que es una mujer de 60 años con un nivel de colesterol de 450 en base a nuestro árbol de clasificación ¿A qué grupo de paciente pertenece?

Podemos observar que en base al árbol de clasificación la paciente pertenece a un grupo de riesgo, es decir es propensa a recibir un ataque cardíaco dado que presenta altos niveles de colesterol.

Para utilizar arboles de decisión en Python se utiliza la librería scikit-learn y la sub-librería DecisionTreeClassifier.

A continuación se muestra la etapa de entrenamiento de un árbol de decisión cuyos elementos de cada nodo son: el factor: "illiteracy" y "birth_mortal", "sample" muestra la cantidad de elementos de la muestra en cada nodo, gini: expresa el coeficiente de Gini del nodo previo la partición, y "value" las cantidades asignadas a cada tipo de variable categórica.



ACTIVIDAD

Considerando la base de datos analizada en la clase sobre indicadores socioeconómicos de la República Argentina, crear una variable categórica llamada "poverty_index" cuyos niveles de pobreza son clasificados como : High, Middle y Low.

1. Siendo los factores: "school dropout", "birth mortal" y "poverty" representar el árbol de clasificación analizando en cada nodo el valor del índice Gini.
2. Representar en una matriz de confusión la performance del modelo y analizar su resultado.