Árboles de Regresión

Roberto Osciel Romero Obispo

La base de datos de este ejemplo está disponible en: UCI Repository

Se crea la clasificación para predecir la variable Y (target), definida como:

- $Y = 1$ si el paciente **sí** sufre una enfermedad cardíaca
- $Y = 0$ si el paciente **no** sufre una enfermedad cardíaca

Descripción de las variables

1) **Edad**: Edad del individuo.
2) **Sexo**: 1 = masculino, 0 = femenino.
3) **Dolor de pecho** (Chest-pain):
   - 1 = angina típica
   - 2 = angina atípica
   - 3 = dolor no anginoso
   - 4 = asintomático.
4) **Presión arterial en reposo**: Muestra el valor de la presión arterial en reposo (mmHg).
5) **Chol**: Colesterol sérico en mg/dl.
6) **Azúcar en sangre en ayunas**: 120 mg/dl. Si el nivel de azúcar es $>$ 120 mg/dl, entonces: 1 = verdadero, de lo contrario: 0 = falso.
7) **ECG en reposo**: Electrocardiograma:
   - 0 = normal
   - 1 = con onda ST-T anormal
   - 2 = hipertrofia ventricular izquierda.
8) **Frecuencia cardíaca**: Máxima alcanzada.
9) **Angina inducida por ejercicio**: 1 = sí, 0 = no.
10) **Depresión del ST inducida por el ejercicio en relación con el reposo**: Valor entero o flotante.

11) **Segmento ST del ejercicio máximo**:
    - 1 = ascendente
    - 2 = plano
    - 3 = descendente.

12) **Número de vasos principales (0-3) coloreados por fluoroscopia**: Muestra el valor entero o flotante.

13) **Thal**: Muestra la talasemia:
    - 3 = normal
    - 6 = defecto fijo
    - 7 = defecto reversible.

Diagnóstico final: Muestra si el individuo sufre o no una enfermedad cardíaca:

- 0 = ausencia
- 1, 2, 3, 4 = presente.

Análisis descriptivo

Demos un primer vistazo a nuestros datos a través de un scatterplot con la variable Edad y el Nivel de colesterol.

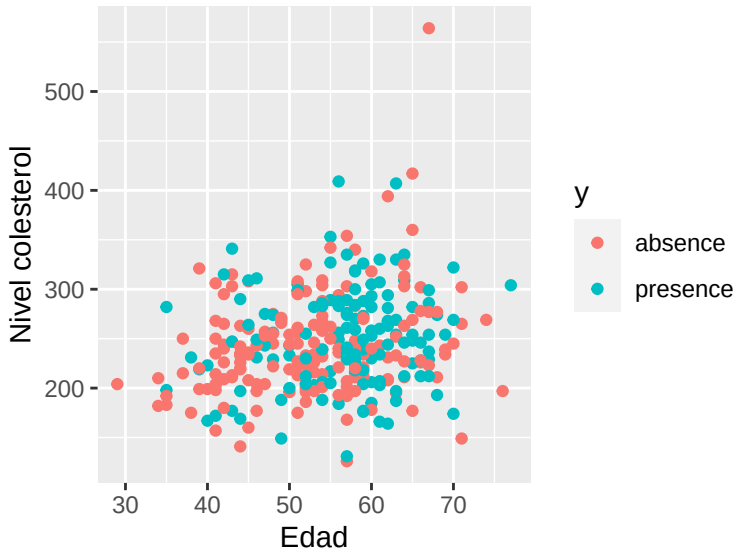

Figura 1: Relación: Edad y Nivel de colesterol

La Figura 1 muestra que ni la edad ni el colesterol parecen ser factores determinantes por sí solos para la presencia o ausencia de enfermedad cardíaca, ya que hay considerable solapamiento entre los grupos.

Modelo Propuesto

Para el análisis, se decidió ajustar un modelo utilizando ahora las variables **Thal**,**ca** y **cp**, utilizando la biblioteca **tree**. Nuestro ajuste fue el siguiente:

```
mod3 <- tree(y ~ cp + ca + thal, data=datos) # Mismo ajuste
```

El árbol correspondiente es el siguiente:

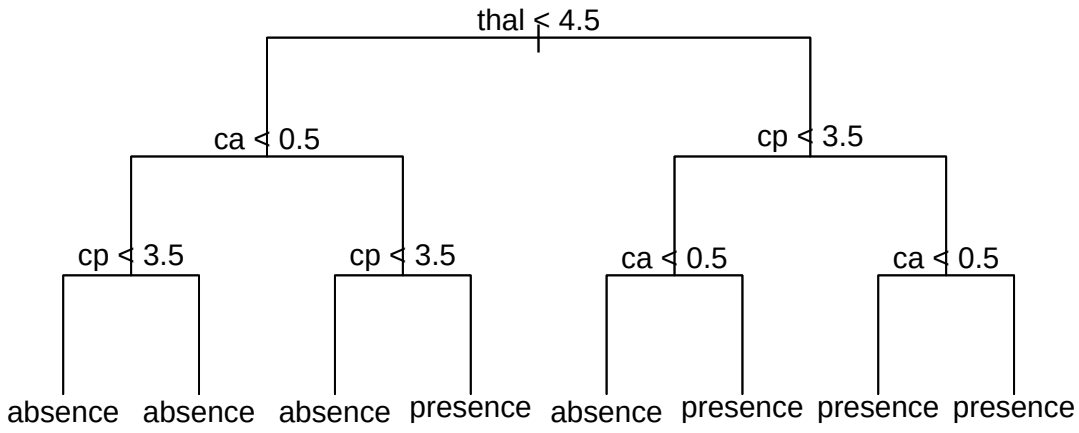

Figura 2: Árbol con Thal, CA y CP como variables

En la Figura 4 tenemos el árbol de decisión está basado en tres variables: thal , ca , y cp , y clasifica la presencia o ausencia de una enfermedad cardíaca.

1. **Nodo raíz**:
   - La primera división se realiza con la variable **thal** (talasemia), específicamente si **thal** < 4.5.
     - Si **thal** < 4.5, seguimos hacia la izquierda.
     - Si **thal** $>= 4.5$, seguimos hacia la derecha.
2. **Rama izquierda** (thal < 4.5):
   - Se hace una segunda división con la variable **ca** (número de vasos coloreados por fluoroscopia), en función de si **ca** < 0.5.
     - Si **ca** < 0.5, seguimos hacia la izquierda.
     - Si **ca** $>= 0.5$, seguimos hacia la derecha.
   - **Subdivisiones basadas en ca** < 0.5:
     - Si **ca** < 0.5, se hace una división adicional basada en la variable **cp** (dolor de pecho), en función de si **cp** < 3.5.
       - Si **cp** < 3.5, se predice **ausencia** de la enfermedad.

        - Si **cp >= 3.5**, se predice **presencia** de la enfermedad.
    - **Subdivisiones basadas en ca >= 0.5**:
        - Si **ca >= 0.5**, se predice **presencia** de la enfermedad cardíaca.
3. **Rama derecha** (thal >= 4.5):
    - Se realiza la misma serie de divisiones que en la rama izquierda:
        - Se divide por **cp < 3.5**, y dependiendo del valor de **ca**, se predice **ausencia** o **presencia** de la enfermedad.
    - **Subdivisiones basadas en cp < 3.5**:
        - Si **cp < 3.5**, se hace una división con **ca < 0.5**:
            - Si **ca < 0.5**, se predice **ausencia** de la enfermedad.
            - Si **ca >= 0.5**, se predice **presencia** de la enfermedad.
    - **Subdivisiones basadas en cp >= 3.5**:
        - Si **cp >= 3.5**, se predice **presencia** de la enfermedad.

En términos generales, las personas con *thal* < 4.5 o *cp* >= 3.5 y ciertos valores bajos de **ca** tienden a no tener la enfermedad, mientras que combinaciones de valores altos de **ca** y **cp** están más asociadas con la **presencia** de la enfermedad cardíaca.

Poder predictivo

Comprobemos ahora el poder predictivo de este nuevo modelo:

```
y_rpart <- predict(mod3, type='class')
tabla3 <- table(datos$y, y_rpart)
tabla3
```

```
##           y_rpart
##            absence presence
##   absence      147       17
##   presence      29      110
```

Se observa que las predicciones correctas superan a las incorrectas, ya que solo hay 29 casos erróneos en las predicciones de **ausencia** y 17 casos erróneos en las predicciones de **presencia**.

```
sum(diag(tabla3)) / sum(tabla3)
```

```
## [1] 0.8481848
```

Observamos que la capacidad predictiva de nuestro modelo, ajustado con las variables **Thal**, **CA** y **CP**, para determinar si un paciente padece o no una enfermedad cardíaca ahora es del 84.81 %. Además obtuvimos un modelo con una interpretación más fácil.

Conclusión

Dado el significado que tienen las 3 variables ocupadas para el modelo propuesto:

- **ca**: probablemente tendría la mayor influencia, ya que indica una obstrucción o daño directo en los vasos sanguíneos, lo que es un predictor claro de enfermedad cardíaca.
- **thal**: también podría ser un buen predictor, ya que la talasemia tiene efectos en el transporte de oxígeno y el esfuerzo cardíaco.
- **cp**: aunque es un síntoma subjetivo, sigue siendo una señal importante para evaluar la presencia de una enfermedad cardíaca, y sus diferentes categorías ayudan a determinar el nivel de riesgo.

Si eligimos un modelo basado solo en estas tres variables, el árbol de regresión utilizaría los valores de **ca** y **thal** como predictores clave, y el **cp** como un indicador de síntomas, lo que podría dar una buena aproximación al riesgo de enfermedad cardíaca.