# Diabetes

Corona López José Luis

2024-12-04

```
library(readxl)
library(dplyr)
library(caret)
library(rpart)
library(rpart.plot)
Base_1 <- read_excel("~/Downloads/Base 1.xlsx")
Base_1 <- Base_1[,-7]
```

## Árbol de decisión

### Preparación

Preparamos la base de datos 1. Eliminar valores faltantes 2. Convertir variables categóricas a factor

```
glimpse(Base_1)
```

```
## Rows: 308
## Columns: 8
## $ Pregnancies   <dbl> 9, 1, 8, 5, 10, 0, 0, 0, 8, 6, 1, 0, 0, 7, 4, 0, 2, 7, 8~
## $ Glucose       <dbl> 120, 71, 74, 88, 115, 124, 74, 97, 120, 154, 144, 137, 1~
## $ BloodPressure <dbl> 72, 62, 70, 78, 98, 56, 52, 64, 0, 78, 82, 70, 66, 90, 6~
## $ SkinThickness <dbl> 22, 0, 40, 30, 0, 13, 10, 36, 0, 41, 40, 38, 27, 0, 0, 2~
## $ Insulin       <dbl> 56, 0, 49, 0, 0, 105, 36, 100, 0, 140, 0, 0, 0, 0, 0, 0,~
## $ BMI           <dbl> 20.8, 21.8, 35.3, 27.6, 24.0, 21.8, 27.8, 36.8, 30.0, 46~
## $ Age           <dbl> 48, 26, 39, 37, 34, 21, 22, 25, 38, 27, 28, 22, 22, 50, ~
## $ Outcome       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,~
```

```
# Convertimos a variables factor
Base_1$Outcome <- factor(Base_1$Outcome, levels = c(0,1), labels = c("negativo","positivo"))

# Eliminamos filas con valores faltantes
Base_1 <- na.omit(Base_1)
```

Dividimos la base de datos: Trainig (80%) and Testing(20%)

```
set.seed(789)
indices <- createDataPartition(Base_1$Outcome, p = 0.8, list = F)
training <- Base_1[indices,]
```

```r
testing <- Base_1[-indices,]

# Verificamos la proporción died and survived
prop.table(table(training$Outcome))
```

```
##
##  negativo  positivo
## 0.6963563 0.3036437
```

```r
prop.table(table(testing$Outcome))
```

```
##
## negativo positivo
## 0.704918 0.295082
```

Proporciones en training:

- Negativo: 69.64%
- Positivo: 30.36%

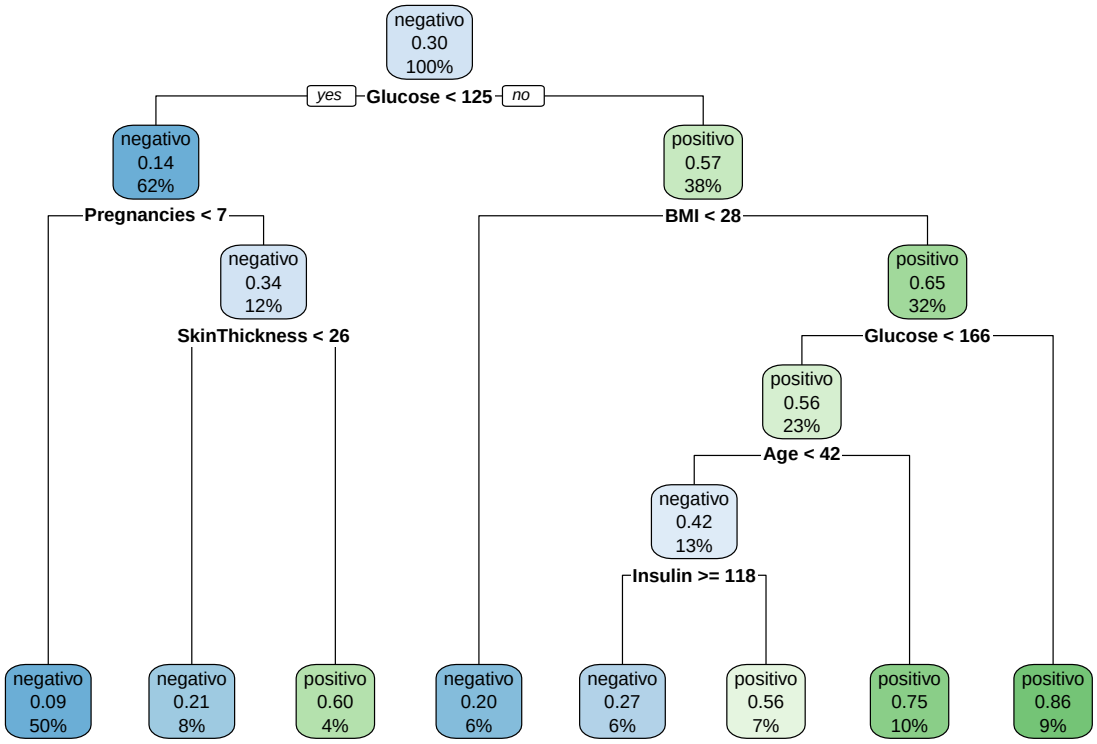
Proporciones en testing:

- Negativo: 70.49%
- Positivo: 29.51%

Construimos el árbol de decisión

## Desarrollo

```r
modelo <- rpart(Outcome ~ ., data = training)
rpart.plot(modelo)
```




El árbol de decisión identifica a la glucosa como el predictor principal de diabetes, estableciendo un punto de corte crítico en 125 mg/dL. Los pacientes con niveles inferiores a este umbral presentan un riesgo bajo (14% de probabilidad), mientras que aquellos con glucosa elevada tienen un riesgo significativamente mayor (57% de probabilidad). Para este último grupo, el modelo incorpora variables secundarias como número de embarazos, espesor cutáneo, IMC y edad, permitiendo una estratificación más detallada del riesgo.

La estructura del árbol revela interacciones clínicamente relevantes, identificando subgrupos de muy alto riesgo (hasta 86% de probabilidad) mediante la combinación de múltiples factores. Esta aproximación proporciona un marco interpretable para la toma de decisiones clínicas, aunque la presencia de nodos con pocas observaciones sugiere la necesidad de validación adicional para confirmar la robustez del modelo en poblaciones más amplias.

Evaluamos el modelo con el testing data

```
forecasted <- predict(modelo, testing, type = "class")
testing$Forecasted <- forecasted


#Generamos la matriz de confusiones
confu <- table(testing$Outcome, testing$Forecasted)
confu
```

```
##
##            negativo positivo
##   negativo       40        3
##   positivo        8       10
```

```r
# Tasa de éxito del modelo
exito <- sum(diag(confu)) / nrow(testing)
print(exito)
```

```
## [1] 0.8196721
```

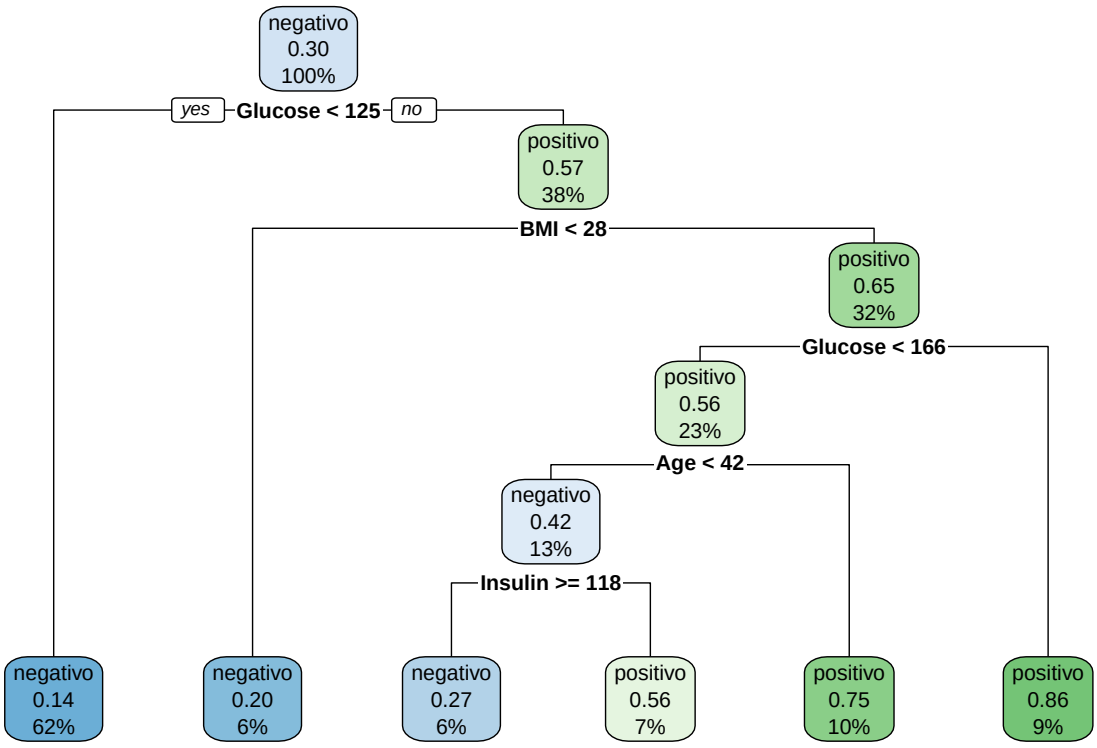
```r
error <- 1 - exito
print(error)
```

```
## [1] 0.1803279
```

El modelo de árbol de decisión demostró un rendimiento satisfactorio en la predicción de diabetes, alcanzando una tasa de éxito del 81.97% en el conjunto de prueba. La matriz de confusión revela que el modelo clasificó correctamente 40 casos negativos y 10 positivos, aunque presentó cierta dificultad para identificar todos los casos diabéticos, con 8 falsos negativos que podrían representar un riesgo clínico significativo. El error global del 18.03% sugiere que, si bien el modelo es útil para la estratificación inicial, podría beneficiarse de ajustes adicionales para mejorar su sensibilidad en la detección de casos positivos.

```r
# Ajustar los parámetros de complejidad del árbol
modelo1 <- rpart(Outcome ~ ., data = training, method = "class", control = rpart.control(cp = 0.02))
#modelo1 <- rpart(Outcome ~ . - DiabetesPedigreeFunction, data = training, method = "class", control = 

# Graficar el árbol ajustados
rpart.plot(modelo1)
```



El árbol podado mantiene la glucosa como variable principal con el mismo punto de corte en 125 mg/dL, pero presenta una estructura notablemente más simple y clínicamente más interpretable. La poda eliminó las divisiones menos relevantes, conservando solo los nodos con mayor poder predictivo: IMC, edad e insulina como variables secundarias. Este árbol simplificado identifica claramente el perfil de máximo riesgo (86% de probabilidad) en pacientes con glucosa elevada, IMC $> 28$ y mayores de 42 años, mientras que el modelo general muestra un buen equilibrio entre complejidad y capacidad predictiva, manteniendo una tasa de acierto del 82% con reglas más robustas y generalizables.

```
forecasted1 <- predict(modelo1, testing, type = "class")
testing$Forecasted1 <- forecasted1


#Generamos la matriz de confusiones
confu1 <- table(testing$Outcome, testing$Forecasted1)
confu1
```

```
##
##            negativo positivo
##   negativo       41        2
##   positivo        9        9
```

```
# Tasa de éxito del modelo
exito <- sum(diag(confu1)) / nrow(testing)
print(exito)
```

```
## [1] 0.8196721
```

```
error <- 1 - exito
print(error)
```

```
## [1] 0.1803279
```

El árbol podado mantiene idéntica efectividad global (81.97% de acierto) que el modelo original, pero presenta un comportamiento diferencial en las clasificaciones: mejora ligeramente en la identificación de casos negativos (aumentando de 40 a 41 aciertos) aunque empeora marginalmente en positivos (disminuyendo de 10 a 9 aciertos). Esta simplificación estructural no compromete la capacidad predictiva general del modelo, pero sugiere un trade-off entre especificidad y sensibilidad, donde la reducción de complejidad parece beneficiar más la identificación correcta de pacientes sanos que de aquellos con diabetes.

## Conclusión

El desarrollo de este modelo de árbol de decisión para predicción de diabetes demuestra la efectividad de los algoritmos de clasificación en el ámbito clínico, alcanzando una precisión global del 82% mediante un modelo interpretable y clínicamente relevante. La identificación de la glucosa como variable principal con punto de corte en 125 mg/dL, complementada con factores como IMC, edad e historial de embarazos, proporciona un marco decisiona sólido para la estratificación de riesgo. La poda del árbol logró simplificar la complejidad del modelo manteniendo su capacidad predictiva, aunque se evidencia el desafío típico en el balance sensibilidad-especificidad, con mejor desempeño en identificar casos negativos que positivos. El proyecto valida la utilidad de los árboles de decisión como herramienta accesible y clínicamente interpretable para el apoyo en diagnóstico, ofreciendo tanto valor predictivo como insights aplicables en la práctica médica real.