Ayudantía 03 Julio

Diplomado en Data Science

Natalie Julian - Paula Muñoz

Ejercicio 1

La base de datos `heart.csv` contiene información de distintos pacientes y sus resultados de distintos exámenes cardiacos. Algunas de las variables de esta base están descritas en la siguiente tabla:

Variable	Descripción
`age`	Edad en años
`sex`	Sexo de la persona (1=Hombre, 0=Mujer)
`cp`	Tipo de dolor en el pecho (1= Angina típica, 2= Angina atípica, 3=No anginoso, 0=Asintomático)
`fbs`	Glicemia en ayunas (1=Mayor a 120 mg/dl, 0=Menor a 120 mg/dl)
`thall`	Presencia de defecto (1=Fijo, 2=Normal, 3=Reversible)
...	...

Ejercicio 1

Considerando a los pacientes de esta base de datos y las variables de la tabla realice los siguientes pasos:

1.- Separe la base de datos en dos grupos, 70% de los datos para generar una base de entrenamiento y el esto para la base de validación.

2.- Ajuste un modelo para predecir el sexo de los pacientes a partir del tipo de dolor en el pecho presentado, la medición de la glicemia en ayunas y la presencia de un defecto.

3.- Indique si el predictor es factor de riesgo, factor protector o no presenta efecto.

4.- Genere la matriz de confusión para el modelo con un punto de corte de 0.7.

5.- Calcule la sensibilidad, la especificidad y la exactitud del modelo.

Ejercicio 1

6.- Calcule el estadístico F_1.

7.- Analice la bondad del ajuste.

8.- Analice el poder predictivo del modelo.

Ejercicio 2

El paquete `ISLR` posee la base de datos `Default` que contiene información de la tarjeta de crédito de distintos clientes. Las variables de esta base están descritas en la siguiente tabla:

Variable	Descripción
`default`	El cliente incumplió el pago de su cuota (1 = Sí, 0 = No)
`student`	El cliente corresponde a un estudiante (1= Sí, 0 = No)
`balance`	Saldo promedio que le queda al cliente en su tarjeta después de realizar el pago mensual
`income`	Ingreso del cliente

Ejercicio 2

a) Analice la asociación/relación de la variable respuesta default con las variables:

- Student
- Balance
- Income

b) Realice un gráfico que muestre los puntos y el comportamiento de una regresión lineal usando como predictor a Balance. Realice lo mismo pero con un modelo de regresión logística. Explique por qué no sería correcto utilizar una regresión lineal, explique también la idea de la regresión logística. Comente.

c) Obtenga set de entrenamiento y prueba al 70%-30% respectivamente. Luego utilice backward para establecer el modelo final. Comente cuál sería la variable que menos aporta al modelo.

d) Interprete los coeficientes del modelo anterior. ¿Qué variables se relacionan con una mayor probabilidad de incumplimiento?

Ejercicio 2

e) Con la función optimalCutoff() del paquete *InformationValue*, determine cuál es el punto de corte óptimo para predecir si un cliente presentó incumplimiento o no. Evalúe la bondad de ajuste del modelo con el punto de corte encontrado, obtenga:

- Curva ROC: ¿En base al área de la curva ROC, usted diría que el modelo es bueno?
- Matriz de confusión: Determine VP, VN, FN, FP y comente qué estaría pasando en este modelo.
- Sensibilidad
- Especificidad
- Precisión