Ejercicio Repaso

**Regresión Logística**

1. Ajustar un modelo de regresión para el conjunto de datos Adult. En el archivo adjunto “Documentacion\_conjunto\_de\_datos”, se encontrará la información con el contenido original del conjunto de datos. El mismo recibió algunas transformaciones que se presentan con código SQL en el final de este documento. El archivo contiene 32.561 registros con información de trabajadores en los EEUU, donde se pretende determinar por sus condiciones socio demográficas si obtuvieron en el futuro un ingreso mayor a 50.000 dólares anuales. Los trabajadores que llegaron a ese importe o lo superaron fueron marcados con 1 y el resto con 0 para determinar el objetivo a analizar.

Se necesita construir un modelo que según las condiciones socio demográficas estime que probabilidad posee de superar ese importe en el futuro.

* + 1. Separe las poblaciones en entrenamiento y validación en forma aleatoria. Indique que cantidad de casos quedaron para cada ambiente.
    2. Formule el mejor modelo posible de regresión logística de acuerdo a la consigna y preséntelo. Se debe entregar el modelo (con la asignación de las categorías originales correspondientes que el software entregó para cada una de las variables dummy)
    3. Calcular el AUC (Curva ROC) en entrenamiento y validación indicando el total de casos de cada una de las clases.
    4. Selecciones el 10% de los trabajadores en el ambiente de validación de acuerdo a la siguiente lógica. Entregue los resultados indicados:
* Al azar e indique la cantidad de trabajadores que llegaron a superar los 50.000 dólares anuales.
* Utilizando el modelo desarrollado en el punto 1.1.2 e indique la cantidad de trabajadores que llegaron a superar los 50.000 dólares anuales.
  + 1. Calcular y/o obtener los siguientes resultados:
* Indicar en cuanto sería el impacto en modificar una unidad de por lo menos una variable continua del modelo.
* Indicar si hay puntos incluyentes con COOK.
* Indicar que método de selección de variables se utilizó y explicar su funcionamiento.
* Mostrar el estadístico de Hosmer-Lemeshow en el último paso del modelo.

////