Ejercicios para practicar: logística binaria II

El conjunto de datos "DatosGSS" contiene información sobre algunas cuestiones sociodemográficas de 2035 encuestados por el *GSSSurvey* (el equivalente estadounidense al CIS). Las variables contenidas en el fichero son:

Variable	Descripción	Codificación
ClaseSocial	Clase social autopercibida	
Felicidad	Respuesta a la pregunta: "¿Cómo de feliz se siente usted?"	Not too happy = No muy feliz Pretty happy = Bastante feliz Very happy = Muy feliz
Politica	Identificación política	
Tamano	Tamaño (en miles de habitantes) del municipio de residencia	
Region	Región de residencia	
Ingreso	Ingresos familiares anuales	0 = Ninguno; 1 = < 25.000$; 2 = ≥ 25.000$
Raza		
Genero		
Edad		
Hijos Var. Depend.	¿El encuestado tiene hijos?	0 = No; 1 = Sí
EstadoCivil		
Empleo	Ocupación actual	
Zodiaco	Signo del zodiaco (12 niveles)	

Los ejercicios consisten en lo siguiente (recuerda que debes llevarlos a cabo sobre un archivo qmd para posteriormente generar el html y subirlo al campus virtual):

1. Carga los datos en el entorno de *Rstudio* a través de la función *readRDS* (nota: como estás trabajando con datos "conocidos" no es necesario revisar el tipo de las variables, pues lo hiciste el primer día, pero se debe hacer siempre que el conjunto de datos es nuevo). Formatea la variable dependiente si fuera necesario.
2. Realiza una partición del conjunto de datos en entrenamiento (80%) y prueba (20%).
3. Utilizando los datos de la partición de entrenamiento, construye 6 modelos de regresión logística binaria para la variable Hijos aplicando los 3 métodos de selección de variables estudiados y los 2 criterios de selección a partir de la función *step*. ¿Cuántos parámetros tienen los modelos? ¿Cuántos modelos diferentes se generan? ¿A qué se deben las diferencias (sobre todo entre los modelos generados con AIC y los generados con BIC)?
4. Una vez generados los modelos (los de la selección automática y el que creaste manualmente con las variables EstadoCivil, Edad y Raza), se debe determinar cuál es el mejor de todos ellos, para lo cual debes aplicar validación cruzada repetida (utiliza un bucle para simplificar esta tarea).
   Genera los boxplots para las 3 medidas (AUC, tasa de acierto e índice Kappa), así como los resúmenes y compara los modelos. ¿Cuál parece ser el mejor? Recuerda que si varios modelos son parecidos en cuanto a capacidad predictiva se debe escoger el más sencillo (el que tenga menos parámetros).
5. Obtén la matriz de confusión utilizando los datos de la partición de prueba del modelo "ganador", así como un resumen de sus estadísticos y el AUC. ¿Qué puedes decir de la calidad del modelo?
6. Para finalizar, realiza un análisis de tipo II sobre el modelo, saca las conclusiones oportunas e interpreta los *odds-ratio*.
7. Construye un modelo LASSO y otro Ridge con todos los valores posibles de λ y representa gráficamente los resultados. Comenta lo que observes y, basándote en dicha información, explica los fundamentos de ambos modelos incidiendo es sus diferencias.
8. Para determinar el modelo penalizado óptimo, aplica validación cruzada para una combinación amplia de α y λ y representa gráficamente los resultados. Indica qué significa la línea horizontal que se añade en el gráfico y comenta los resultados que observas.
   Determina la combinación de parámetros óptima según la regla *1se* y, posteriormente, construye el modelo definitivo. Obtén los parámetros del modelo y comenta su efecto sobre la variable objetivo.
9. Obtén la matriz de confusión utilizando los datos de la partición de prueba del modelo anterior, así como un resumen de sus estadísticos. Calcula también su AUC. ¿Qué puedes decir de la calidad del modelo?