1. El paquete ISLR contiene los datos **Default** que corresponden a datos de cesación de pago de clientes de tarjetas de crédito. Las variables registradas son: **default** (yes o no), **student** (yes o no), **balance** (balance mensual de la tarjeta) e **income** (ingreso anual). Se desea predecir si un individuo tendrá un default en su tarjeta de crédito o no con las variable **balance** y/o **income**.

   a) Realizar un plot de las variables **balance** vs. **income** coloreando de rojo los puntos que corresponden a default="Yes" y azul con default= "No'. ¿Qué se observa en este gráfico? ¿Da la impresión de que las dos variables tienen la misma capacidad predictiva?

   b) Realizar boxplots paralelos para la variable **balance** clasificando por la variable **default**. ¿Qué se observa? Repetir con la variable **income**. ¿Le parece que la asociación entre las dos variables con **default** es la misma?

   c) Implementar una función que **vecinos.k** que tenga por argumento un vector de datos $\mathbf{x} = (x_1, \ldots, x_n)$, un valor k y un punto x y que devuelva los índices de los k vecinos más cercanos al punto x.

   d) Implementar una función **clasificador** que clasifique a un punto x de acuerdo a la regla de clasificación de $k-$vecinos más cercanos usando como input de variable clasificadora a $\mathbf{x} = (x_1, \ldots, x_n)$ y como etiquetas a $\mathbf{y} = (y_1, \ldots, y_n)$.

   e) Aplicar la regla implementada en el ítem anterior al conjunto de datos **Default** para clasificar un punto con valor de balance $x = 1500$, utilizando como input de la variable clasificadora los datos de **balance** y como input de etiquetas a los datos de **default** con $k = 5$. Repita con $x = 1700$ y con $x = 2000$.

   f) Dividir al azar el conjunto de datos del archivo **Default** en dos submuestras: una de tamaño 8000 que llamaremos datos de entrenamiento (datos.E) y otra de tamaño 2000 que llamaremos de testeo (datos.T). Utilizar la semilla set.seed(123) a los fines de poder replicar este experimento numérico.

   g) Considerar ahora la regla **clasificador**, implementada en item anterior, cuando utiliza como input de la variable clasificadora los datos de **balance** de datos.E y como input de etiquetas los datos de **default** datos.E, con $k = 5$. Utilizar ahora esta regla para clasificar los puntos en datos.T, y calcule el porcentaje de observaciones mal clasificadas.

h) Repetir el ítem anterior usando como input de la variable clasificadora a **income**. Comparar con el resultado del ítem anterior.

i) El comando glm permite ajustar una regresión logística a un conjunto de datos. Así, para estimar los parámetros de una regresión logística utilizando todos los datos del archivo **Default** por el método de máxima verosimilitud para predecir la probabilidad de que un cliente tenga un default a partir de la variable **balance** haríamos

```
glm(default~balance,family=binomial)
```

mientras que

```
glm(default~balance,family=binomial)$fitted
```

da los valores predichos de las probabilidades usando las estimaciones de los parámetros obtenidas por máxima verosimilitud.

Utilizando el modelo de regresión logística, implementar una regla para clasificar a un individuo con un valor x de **balance**.

¿Qué porcentaje de puntos clasifica incorrectamente esta regla en el conjunto de datos.T cuando utiliza como muestra de entrenamiento a datos.E?

j) Repetir el ítem anterior utilizando ahora como variables clasificadoras **balance** e **income** simultáneamente. Comparar los resultados al usar esta regla y las de los ítems anteriores.

2. La cifosis es una deformación en la columna vertebral que puede presentarse en niños con cirugía correctiva de la espina dorsal. Se cree que la incidencia de deformaciones (**kyp**=1 si la deformación está presente, **kyp**=0 caso contrario) después de la cirugía de la columna vertebral está asociada a la edad en la que ésta se realiza (**Age**, en meses), a la primera vértebra operada (**Start**) y al número de vértebras involucradas en la operación (**Number**).

   El conjunto de datos kyphosis.csv contiene la información correspondiente a 81 niños.

   a) Graficar la respuesta binaria de la incidencia de la cifosis versus la edad del niño. Ajustar un modeo logístico simple usando la variable edad como regresora. Examinar el ajuste y la significación de **Age**.

   b) Ajustar un modelo de regresión logística usando como covariables **Age**, **Start** y **Number**. Examinar el ajuste y la significación de todos las componentes del modelo.

c) Ajustar un modelo de regresión logística cómo en el ítem a) pero agregando un término cuadrático en **Age**. Examinar el ajuste y la significación de todos las componentes del modelo.

d) Ajustar un modelo de regresión logística usando como covariables **Age**, **Start**, **Number** y una componente cuadrática en **Age**. Examinar el ajuste y la significación de todos las componentes del modelo. Comprarar con el modelo del ítem anterior.

e) ¿Es más conveniente considerar un modelo que contenga una componente cuadática en Number y una interacción entre **Age** y **Number**? Justificar.