Minería de Datos

Adrián Edreira Gantes

Lucas García García

Sergio Liste Vázquez

Tarea 1

El objetivo de esta tarea es determinar los grupos existentes para el conjunto de datos actual, reduciendo dicho conjunto a un pequeño número de grupos que permita entender mejor los datos originales. Se seleccionarán diferentes variables para realizar dichos grupos y utilizaremos K-means durante el proceso.

Selección de variables

Para dividir en grupos los datos se han seleccionado las siguientes variables:

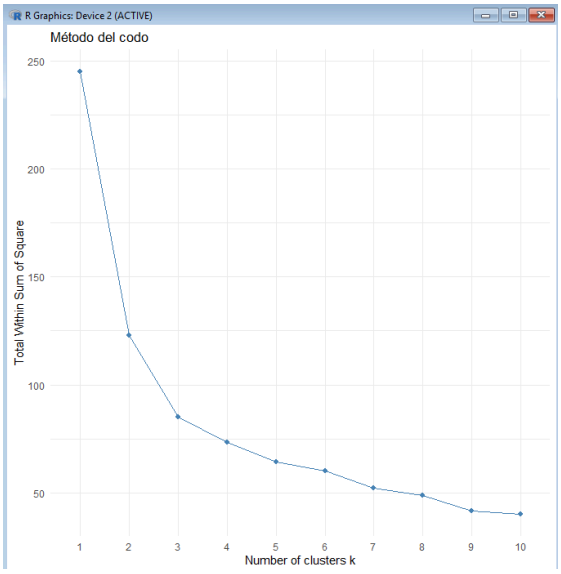
- Porcentaje de graduados: representa el nivel educativo.
- Ingresos per cápita: indica el nivel económico.
- Esperanza de vida: relaciona salud y calidad de vida.
- Tasa de asesinatos: indica zonas con alta criminalidad.
- Analfabetismo: representa zonas de baja educación

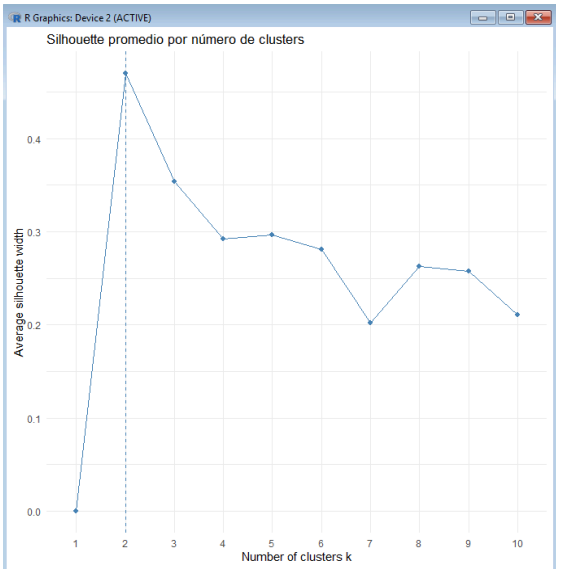
Preparación de datos

Crearemos un conjunto de datos con las variables seleccionadas, lo que nos permitirá entre otras cosas normalizar fácilmente los datos en rattle y obtener el número adecuado de clusters para el conjunto de datos.

Para normalizar los datos usaremos rattle, en la pestaña transformar, y usaremos la opción de escala[0-1].

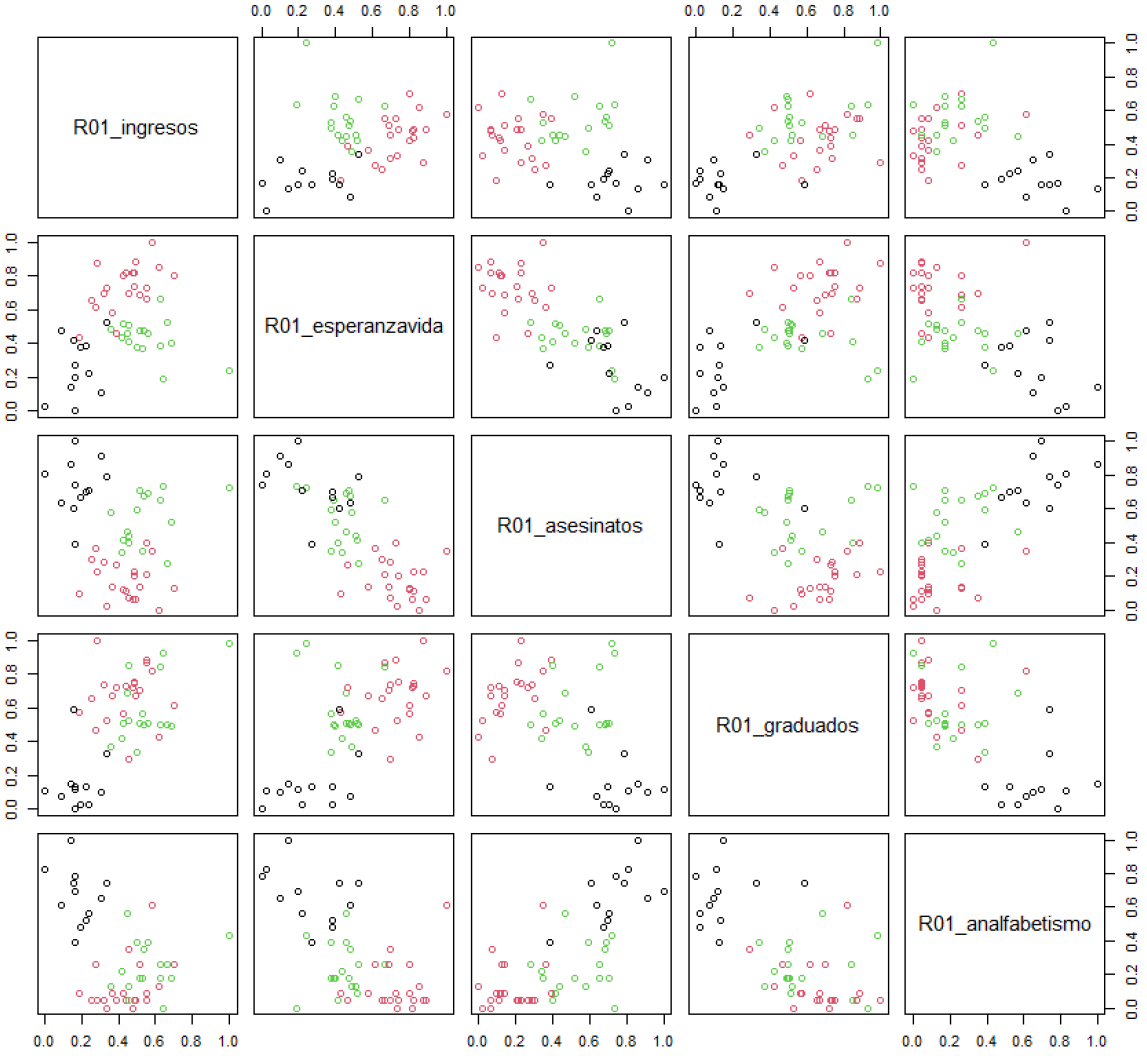
Para obtener el número necesario de clusters utilizaremos los métodos de Elbow (Método del codo) y de Silhouette. Tras observar las gráficas vemos que el número de clusters más equilibrado sería 3 ya que en la primera la reducción de WSS empieza a ser menor con cada cluster adicional y en la segunda es la más alta por detrás de los 2 clusters.



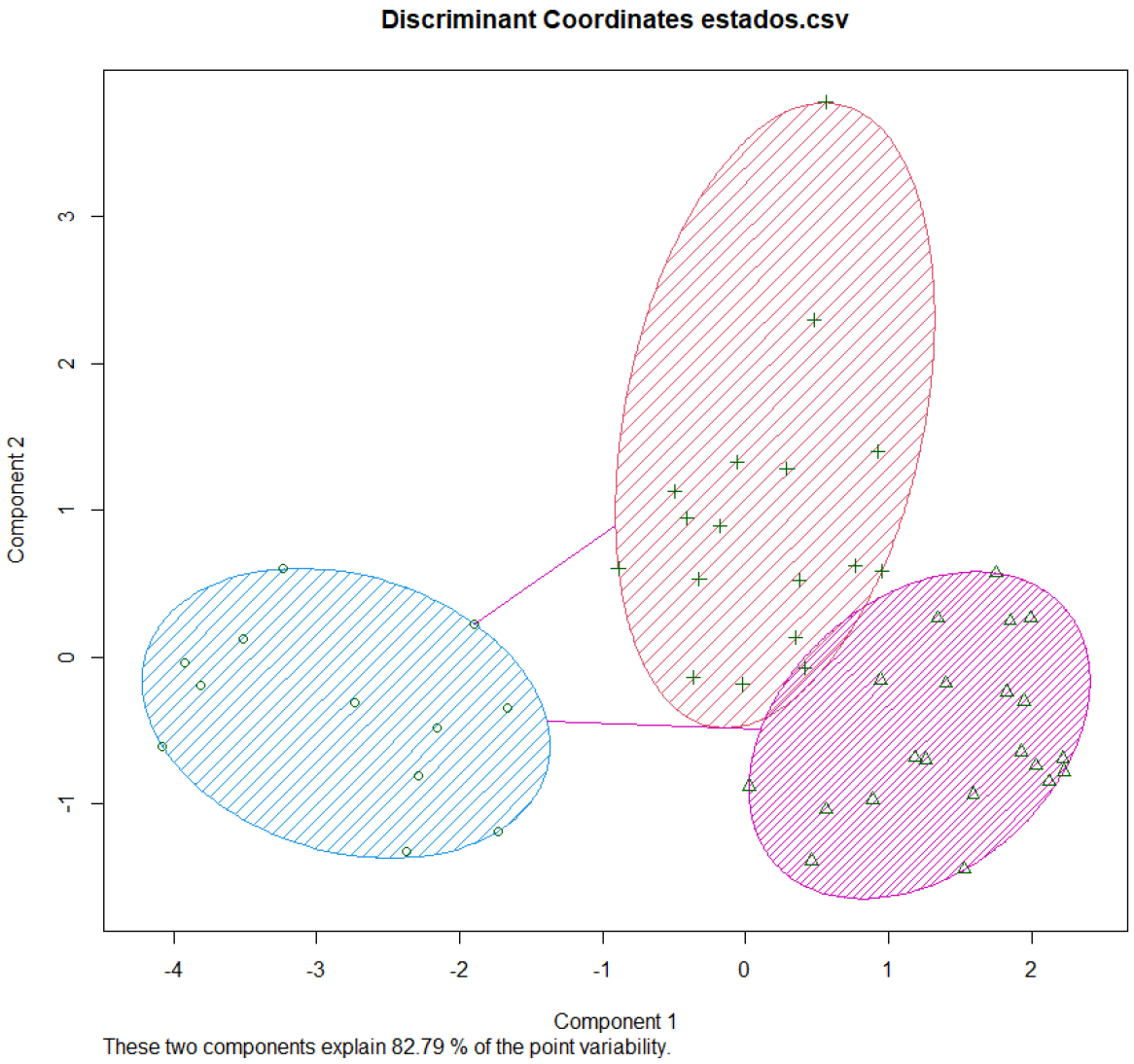


Creación de clusters

Para la creación de clusters usamos la opción de rattle dedicada a esta tarea. Seleccionamos K-means e indicamos el número de clusters calculado anteriormente. Tras ello, obtenemos diferentes resultados y gráficas que podremos visualizar. En esta de aquí observamos como los datos están asociados a un cluster concreto y se les asigna un color determinado por cada cluster al que pertenecen, y vemos que en gran medida hay una división de los datos.



Por otro lado tenemos la siguiente gráfica que nos muestra esa división de los datos en sus correspondientes clusters. En el pie de la gráfica vemos un valor que nos indica el porcentaje en el que las variables seleccionadas explican la variabilidad de los datos, la cual es de un 82.79%.



Interpretación de resultados

```
Centros de clústers:

  R01_ingresos R01_esperanzavida R01_asesinatos R01_graduados R01_analfabetismo
1    0.1815615         0.2609338      0.7354015     0.1474576         0.6666667
2    0.4370976         0.7312563      0.1797011     0.6736077         0.1242236
3    0.5549928         0.4345015      0.5422928     0.5900299         0.2327366
```

Tras estudiar los resultados obtenidos en el proceso vemos que en el primer grupo existe una alta tasa de asesinatos y de analfabetismo, así como una baja tasa de graduados y de ingresos. El segundo cluster vemos que tiene una alta esperanza de vida y un alto porcentaje de graduados, es el caso opuesto al cluster número 1. Por último, el último cluster puede representar la media, el grupo intermedio entre los dos clusters anteriores, ya que tiene valores muy próximos a 0.5 en la mayoría de sus variables.

Conclusión

En esta tarea se aplicó un análisis de clustering con K-means sobre datos de los estados de EE. UU., utilizando variables normalizadas como ingresos, esperanza de vida, asesinatos, graduados y analfabetismo. Se identificaron tres grupos bien diferenciados: uno con peores indicadores sociales, otro con buen desarrollo y uno intermedio. La normalización fue clave para un análisis equilibrado.

Podríamos obtener mejores clusters utilizando diferentes variables, semillas o número de clusters, ya que si nos fijamos en el diagrama de discriminantes de coordenadas podemos ver como los perímetros creados para los clusters interseccionan, lo que podría producir errores para datos que aparezcan en dicha zona.

Tarea 2

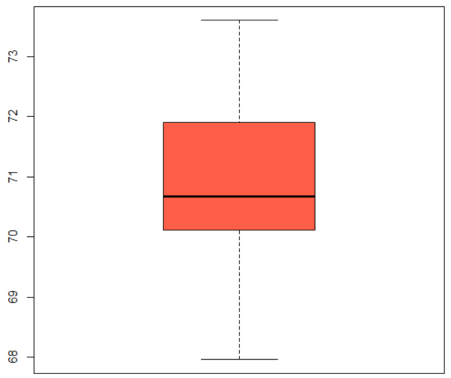
El objetivo de esta tarea es predecir la esperanza de vida de una población concreta utilizando regresión lineal múltiple. Para ello, hemos realizado una serie de pasos para alcanzar dicho objetivo. Comenzamos con la preparación de los datos, realizamos el correspondiente análisis exploratorio, creamos y evaluamos el modelo necesario para, por último, predecir resultados sobre un conjunto de datos.

1. Preparación de datos

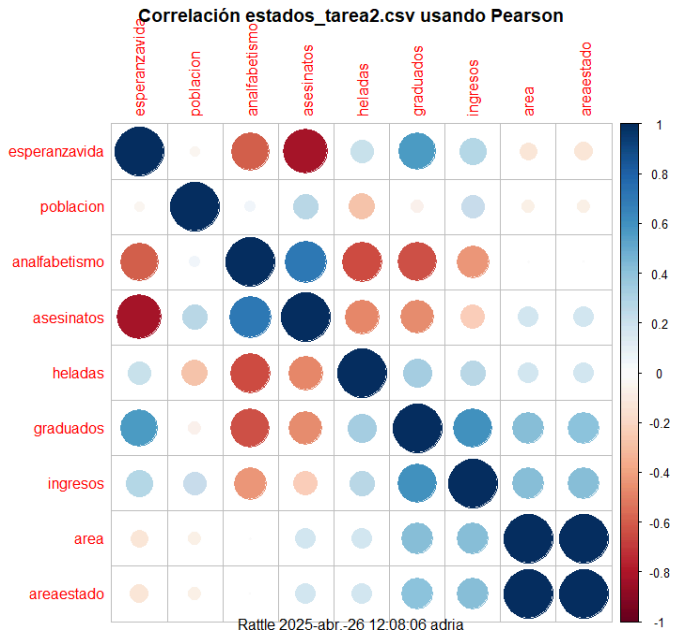
Para limpiar los datos hemos revisado el sumario de los mismos con el objetivo de ver los tipos de valores de las variables y que estén dentro de un rango aceptable. Hemos descartado las columnas “nombre” y “abreviatura” ya que no tienen importancia para el análisis que vamos a realizar, así como separamos la variable objetivo, “esperanzavida”. Por último, al tratarse de un problema de regresión lineal, es preciso crear una variable categórica para la región, ya que representa un código y no un valor numérico que expresa una cantidad.

2. Análisis exploratorio

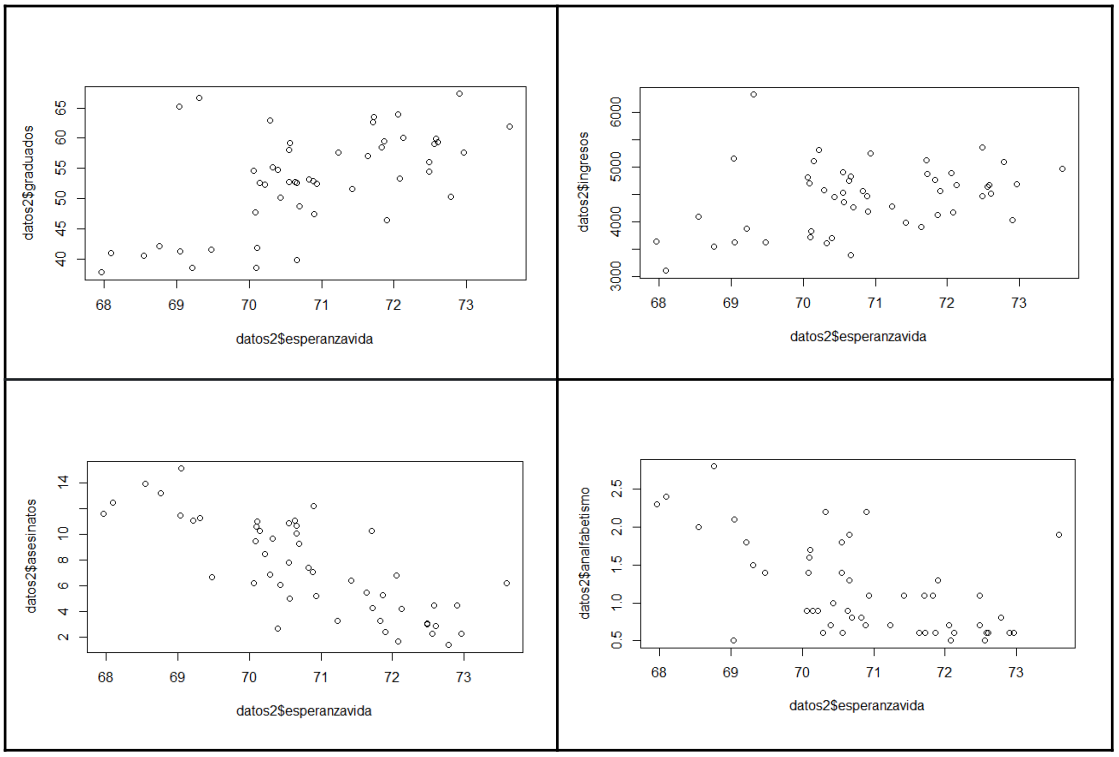
Según el análisis exploratorio vemos que la esperanza de vida está entre los 68 y 74 años, y que la mayoría de datos entre los 70 y 72 años como vemos en el primer y tercer cuartil de la siguiente gráfica, con una media de 70,88 años.



También buscamos si existía alguna correlación entre variables con la variable objetivo. Tras analizar los resultados vemos que variables como el porcentaje de graduados, el número de heladas o la cantidad de ingresos aumentan la esperanza de vida; mientras que otras como el número de asesinatos o el analfabetismo la disminuyen en gran medida. Esto se puede ver perfectamente en la siguiente imágen.



Tras los resultados anteriores decidimos obtener los gráficos correspondientes a las parejas de datos más relacionados con nuestra variable objetivo. Las dos primeras gráficas relacionan la esperanza de vida con el porcentaje de graduados y la media de ingresos, en la que podemos confirmar lo observado en el gráfico anterior, a mayor número de graduados o mayor cantidad de ingresos, mayor esperanza de vida. Las segundas muestran esa disminución de la esperanza de vida, una conforme se producen un mayor número de asesinatos y otra con un mayor porcentaje de analfabetismo.



3. Modelo

Para la creación del modelo de regresión lineal múltiple usaremos la función “lm” de Rattle, usando “esperanzavida” como variable de destino y el resto de variables, salvo las descartadas en el paso 1, como variables de entrada. Usaremos el método de partición entrenamiento/validación/prueba de 70/15/15 y después usaremos diferentes semillas (10, 20, 30, 40 y 50) para obtener el modelo que mejor se comporte.

Empezando con todas las variables obtenemos los siguientes coeficientes para ellas, esto nos ayudará a descartar columnas que no resulten interesantes, como puede ser en este primer caso “areaestado” debido a su p-valor, como podemos observar en la siguiente imagen.

```
Coefficients:
                    Estimate   Std. Error t value Pr(>|t|)
(Intercept)     66.349388020  3.369009337  19.694 6.77e-16 ***
areaestado      -0.000019934  0.000114924  -0.173   0.8638
poblacion        0.000036263  0.000046113   0.786   0.4397
ingresos        -0.000005797  0.000342242  -0.017   0.9866
analfabetismo    0.427988357  0.607574206   0.704   0.4882
asesinatos      -0.235362498  0.093275718  -2.523   0.0190 *
graduados        0.107442228  0.053575150   2.005   0.0568 .
heladas         -0.003525299  0.005188436  -0.679   0.5036
area             0.000017278  0.000119799   0.144   0.8866
IFC_region(1,2]  0.704634291  0.708070519   0.995   0.3300
IFC_region(2,3]  0.841918247  0.562776159   1.496   0.1482
IFC_region(3,4]  0.262410338  0.799932231   0.328   0.7458
```

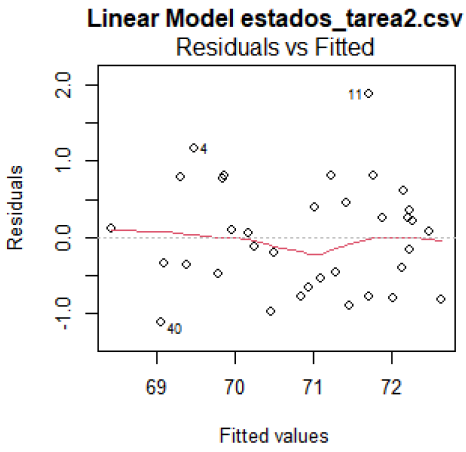
Probando con las diferentes semillas, con diferentes variables y calculando el error de cada modelo, llegamos a la conclusión de que aquel que mejor se comporta es el que se entrena con las variables “poblacion”, “asesinatos”, “analfabetismo”, “ingresos” y “graduados”. Tiene una media de error de 0,575 años respecto a la esperanza de vida y probando con nuevas semillas tiene un comportamiento muy regular. Si nos fijamos en los coeficientes de la siguiente imagen observamos que los mejores predictores para el modelo son “población”, “asesinatos” y “graduados”. Además, vemos que los valores de R^2 y R^2 ajustado son bastante ajustados e indican que el modelo explica una cantidad razonable de la esperanza de vida; así como un p-valor muy pequeño para la estadística F, que sugiere que el modelo es globalmente significativo.

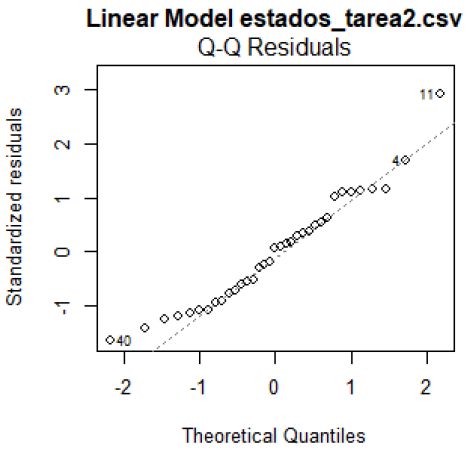
```
Coefficients:
                Estimate  Std. Error t value    Pr(>|t|)
(Intercept)   70.17894415  1.49844391  46.835     < 2e-16 ***
poblacion      0.00009243  0.00005408   1.709      0.0981 .
ingresos      -0.00015932  0.00029560  -0.539      0.5940
analfabetismo  0.45552629  0.32915045   1.384      0.1769
asesinatos    -0.34299805  0.05136462  -6.678 0.000000253 ***
graduados      0.05664871  0.02787439   2.032      0.0514 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.752 on 29 degrees of freedom
Multiple R-squared:  0.7371,    Adjusted R-squared:  0.6917
F-statistic: 16.26 on 5 and 29 DF,  p-value: 0.0000001192
```

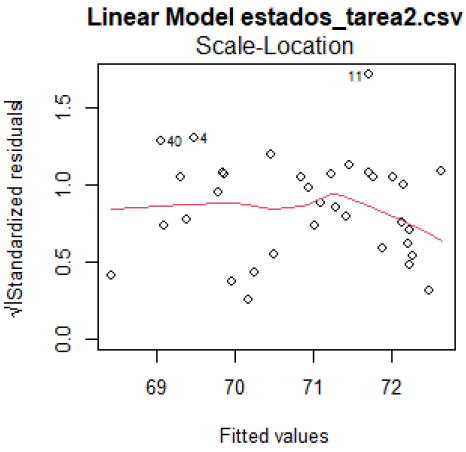
Por otra parte, en el análisis de la varianza vemos que las variables más significativas son “analfabetismo” y “asesinatos”, mientras que el resto de variables afectan en menor medida al resultado del modelo.

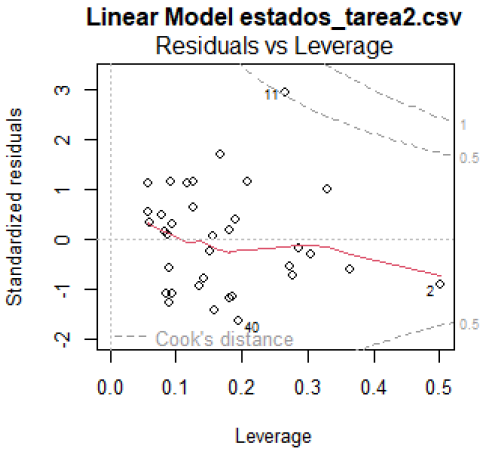
```
Response: esperanzavida
               Df  Sum Sq Mean Sq F value       Pr(>F)
poblacion       1  1.0867  1.0867  1.9218      0.17623
ingresos        1  4.0053  4.0053  7.0830      0.01255 *
analfabetismo   1 12.7185 12.7185 22.4914 0.0000518679 ***
asesinatos      1 25.8212 25.8212 45.6623 0.0000002042 ***
graduados       1  2.3355  2.3355  4.1302      0.05137 .
Residuals      29 16.3990  0.5655
```

Respecto a los residuos podemos ver que nuestro modelo está bastante bien entrenado, sugieren que las predicciones de nuestro modelo están razonablemente cerca de los valores reales, lo podemos ver perfectamente en las siguientes gráficas de los residuos de nuestro modelo.





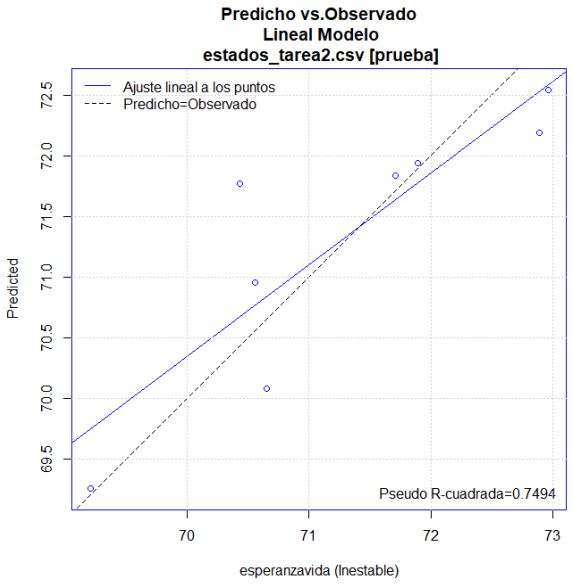




4. Predicción

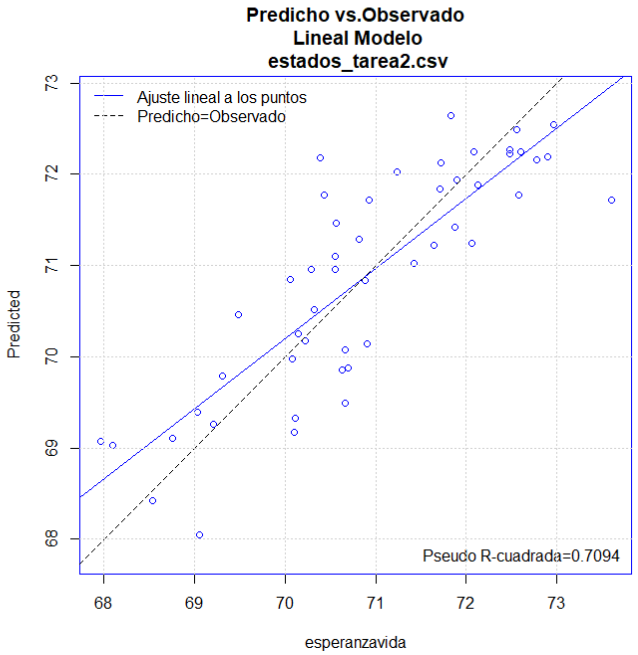
Para la predicción se ha usado el conjunto de datos de prueba para evaluar nuestro modelo. Al ser un conjunto pequeño de datos, el subconjunto de prueba tiene tan solo 8 filas. Usaremos la semilla 50 para esta predicción. Además, realizamos pruebas para distintas semillas y encontramos una gran estabilidad en los modelos creados.

Respecto a la siguiente gráfica, vemos como el modelo responde con gran corrección a la introducción de datos nuevos. Con una pseudo R^2 de casi el 75% podemos afirmar que nuestro modelo predice de forma bastante correcta la esperanza de vida.



Valor real	Predicción	Diferencia
71.71	71.83	-0.13
70.66	70.08	0.58
72.96	72.54	0.42
70.55	70.95	-0.40
69.21	69.26	-0.05
70.43	71.77	-1.34
71.90	71.94	-0.04
72.90	72.19	0.71

Fijándonos en la segunda gráfica, con el conjunto de datos completo, vemos que el modelo se ajusta con buena forma a los datos reales.



5. Conclusión

Tras estudiar los resultados, errores, datos, modelos y gráficas obtenidas llegamos a la conclusión de que las variables que más intervienen en la esperanza de vida son el porcentaje de graduados, la población, y en especial los asesinatos. Además, hemos aprendido a preparar y analizar los datos, así como a estudiar los resultados obtenidos y para acabar obteniendo un modelo bien entrenado mediante regresión lineal múltiple para predecir la esperanza de vida según diferentes factores.

Tarea 3

El objetivo de esta práctica es hacer una tarea de clasificación, que se basa en predecir si un estado de EE.UU. puede considerarse “rico” o “no rico”, en base a variables sociales, educativas y demográficas. Para esto, empleamos la regresión logística, que es una técnica que se adecua para variables binarias como en este caso.

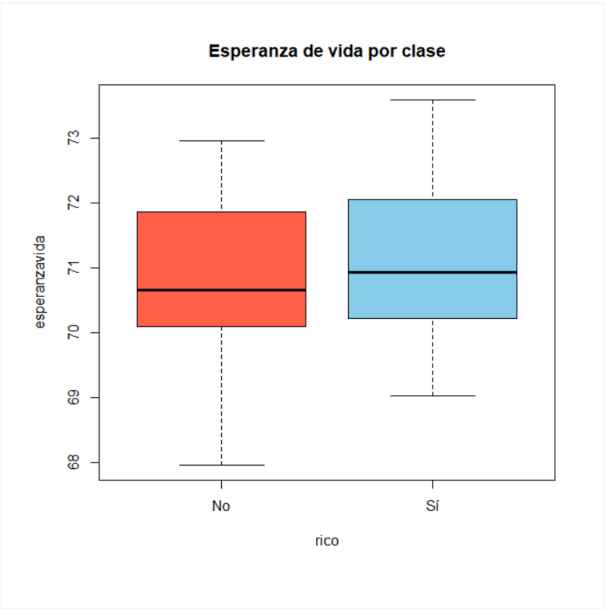
1. Preparación de datos

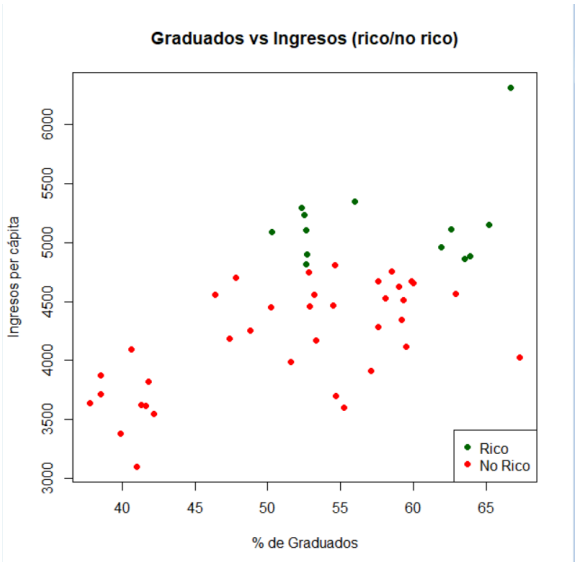
Para preparar los datos, primero se revisó las variables y sus tipos, a continuación eliminamos las columnas nombre, abreviaturas e ingresos, ya que no aportan valor al modelo.
A continuación se creó una variable llamada “rico”, que clasifica a un estado como “rico” si su ingreso per cápita se encuentra por encima del percentil 75 del conjunto de datos.
Ingresos se eliminó porque mantenerla entre las variables predictoras haría que el modelo predijera “rico” de forma trivial, sin utilizar el resto de variables socioeconómicas o demográficas.
También se transformó la variable “región” como categórica, que es más adecuada para su uso en una regresión logística.

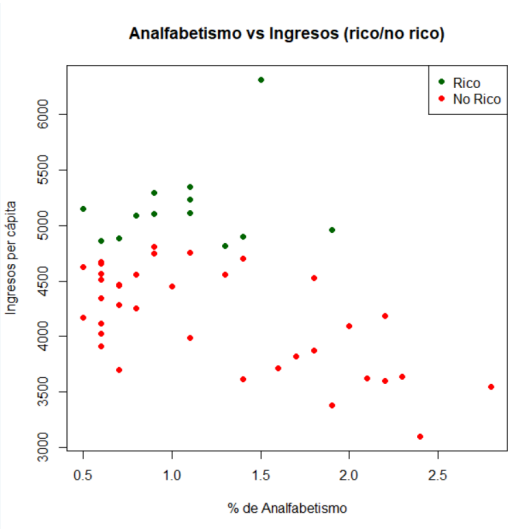
2. Análisis exploratorio

Durante el análisis exploratorio se observaron diferencias notables entre los estados clasificados como ricos y no ricos.
Los estados ricos suelen tener mayor porcentaje de graduados y mayor esperanza de vida, sin embargo los estados no ricos presentan tasas más altas de analfabetismo y asesinatos.

La matriz de correlaciones muestra que variables como “graduados”, “esperanzavida” y “heladas” están positivamente relacionadas con los ingresos, mientras que “analfabetismo” y “asesinatos” tienen una correlación negativa.





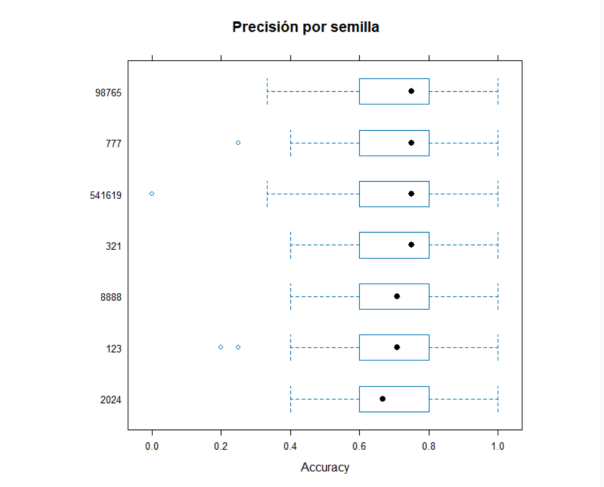


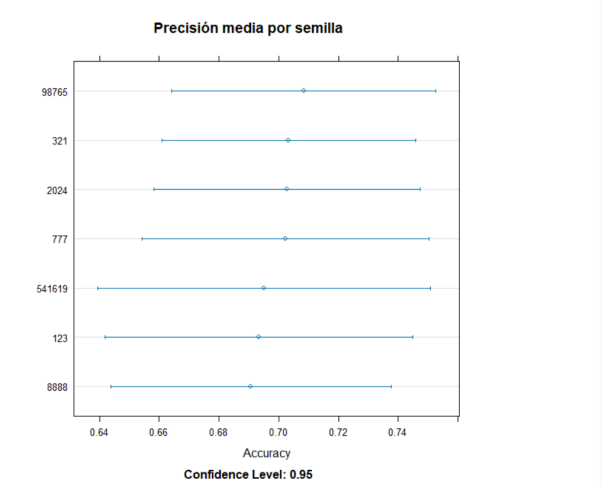
3. Modelo

Para predecir si un estado es rico o no, utilizamos la función train del paquete caret en R, con el método “glm” (modelo lineal generalizado) y familia “binomial”. Esta configuración permite ajustar una regresión bajo un entorno de entrenamiento controlado, integrando técnicas como validación cruzada, pudiendo comparar distintos modelos entrenados con distintas semillas.

La estrategia de evaluación del modelo se centró en la creación de distintas particiones aleatorias del conjunto de datos, utilizando semillas distintas, en combinación con una validación cruzada, para conseguir una medida fiable del modelo.

Se seleccionaron las siguientes semillas: 123, 321, 777, 2024, 8888, 98765 y 541619, con el objetivo de cubrir una muestra diversa sin necesidad de evaluar todas las posibles combinaciones.





Los gráficos generados muestran que las semillas 98765, 321 y 777 proporcionan una precisión media elevada y una menor variabilidad, además de baja dispersión, lo que sugiere que el modelo entrenado con estas semillas tiene un comportamiento más estable y generalizable. Por tanto nos quedaremos con la semilla 321.

4. Predicción

El modelo predice si un estado es “rico” o no y se comparan las predicciones del modelo frente a los valores reales del conjunto de prueba, consiguiendo una precisión del 83.3% lo que indica una buena capacidad para clasificar correctamente los estados en función de su riqueza.

```
Models: 123, 321, 777, 2024, 8888, 98765, 541619
Number of resamples: 50

Accuracy
            Min. 1st Qu.    Median      Mean 3rd Qu. Max. NA's
123    0.2000000     0.6 0.7083333 0.6933333     0.8    1    0
321    0.4000000     0.6 0.7500000 0.7033333     0.8    1    0
777    0.2500000     0.6 0.7500000 0.7023333     0.8    1    0
2024   0.4000000     0.6 0.6666667 0.7026667     0.8    1    0
8888   0.4000000     0.6 0.7083333 0.6906667     0.8    1    0
98765  0.3333333     0.6 0.7500000 0.7083333     0.8    1    0
541619 0.0000000     0.6 0.7500000 0.6950000     0.8    1    0

Kappa
             Min.    1st Qu. Median       Mean   3rd Qu. Max. NA's
123    -0.5000000 -0.2500000      0 0.10811688 0.4464286    1    0
321    -0.3636364 -0.1875000      0 0.09090909 0.2291667    1    0
777    -0.5000000 -0.1875000      0 0.09854978 0.2767857    1    0
2024   -0.3636364 -0.2500000      0 0.11740260 0.2857143    1    0
8888   -0.3636364 -0.2500000      0 0.08168831 0.1875000    1    0
98765  -0.5000000 -0.1153846      0 0.09376290 0.1250000    1    0
541619 -0.5000000 -0.2500000      0 0.10591291 0.2291667    1    0
```

En esta tabla podemos observar que la mayor precisión, nos la dan las semillas 321 y 98765.
Y aunque los valores de kappa son modestos, todos los modelos muestran un nivel de acuerdo leve pero consistente.

5. Conclusión

El modelo de regresión logística entrenado con la semilla 321 ha demostrado un comportamiento robusto, con alta precisión y bajo margen de error. Al ser un modelo estable, es un candidato ideal para generalizar nuestros datos. Así hemos conseguido un modelo bien ajustado y fiable, capaz de clasificar con éxito la riqueza.