Tarea 1 Clasificación

Beltran Aller Lopez

10/11/2019

```
library(rpart)
library(rpart.plot)
library(partykit)
library(party)
library(readxl)
library(fastDummies)
library(ggplot2)
library(MASS)
library(biotools)
library(car)
set.seed(1234)
```

DATA

Disponemos de un dataframe compuesto por un total de 80 observaciones y 9 variables. Estas son las siguientes:

- TIPO: variables categórica, representa el perfil de riesgo en el cual se encuentra enmarcado el cliente. Puede adoptar 3 posibles valores, alto riesgo, riesgo medio y bajo riesgo.
- I: variable numérica, indica la cantidad de ingresos anuales del cliente, está medida en miles de euros.
- Edad: varibale numérica, edad del cliente en cuestión.
- Sexo: variable categórica, el cliente se puede tratar de un hombre o una mujer.
- EC: variable categórica, estado civil del cliente, soltero o casado.
- H: variable numérica, representa el número de hijos del cliente.
- P: variable numérica, patrimonio en miles de euros del cliente.
- R: variable numérica, ratio de endeudamiento del cliente sobre el patrimonio.
- A: variable categórica,grado de aversión al riesgo del cliente, puede tomar 3 posibles valores, alto, medio y bajo.

En la siguiente tabla se muestra un resumen de las características más relevantes de cada variable.

```
##      TIPO                I              Edad          Sexo
##  Length:80          Min.   : 9.30   Min.   :19.00   Length:80
##  Class :character   1st Qu.:15.45   1st Qu.:27.00   Class :character
##  Mode  :character   Median :18.30   Median :31.00   Mode  :character
##                     Mean   :17.65   Mean   :32.31
##                     3rd Qu.:19.73   3rd Qu.:37.25
##                     Max.   :25.30   Max.   :52.00
##       EC                 H               P                R
##  Length:80          Min.   :0.00   Min.   : 32.00   Min.   :20.00
##  Class :character   1st Qu.:0.00   1st Qu.: 53.00   1st Qu.:36.00
##  Mode  :character   Median :1.00   Median : 58.00   Median :47.00
##                     Mean   :1.05   Mean   : 59.73   Mean   :46.86
##                     3rd Qu.:2.00   3rd Qu.: 66.00   3rd Qu.:56.00
##                     Max.   :4.00   Max.   :102.00   Max.   :68.00
```

```
##       A
##  Length:80
##  Class :character
##  Mode  :character
##
##
##
```

ANALISIS DESCRIPTIVO

Acorde a la información visualizada en el dataset, de cara a realizar un análisis discriminante procedo a eliminar las variables categóricas. Por tanto eliminaré el sexo, el estado civil y el grado de aversión al riesgo.

Divido la muestra en función del riesgo de cada una de las observaciones, por tanto genero 3 dataframes distintos, cada uno correspondiente a un perfil de riesgo distinto. Llevaré a cabo un análisis sobre cada una de las variables asociadas a los distintos tipos de riesgo.

CONDICIONES DEL ANALISIS DISCRIMINANTE

- Cada predictor que forma parte del modelo se distribuye de forma normal en cada una de las clases de la variable respuesta. En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases.
- La varianza del predictor es igual en todas las clases de la variable respuesta. En el caso de múltiples predictores, la matriz de covarianza es igual en todas las clases. Si esto no se cumple se recurre a Análisis Discriminante Cuadrático (QDA).

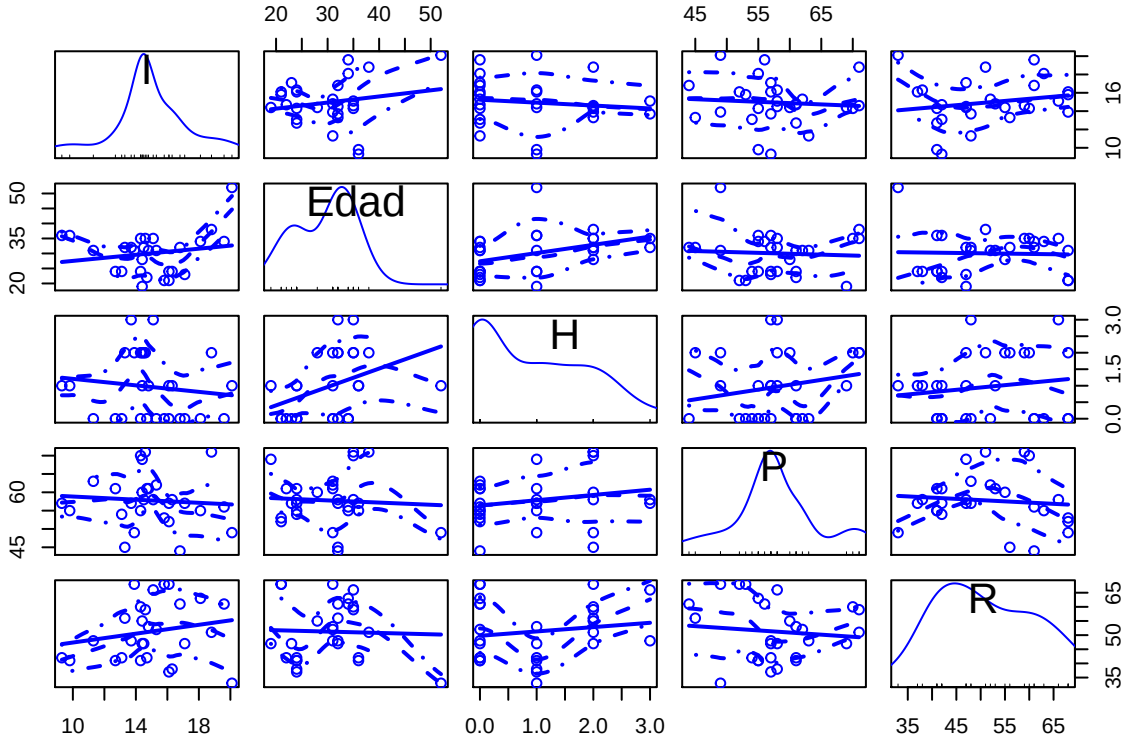
RIESGO ALTO

Del conjunto de las 80 observaciones correspondientes al dataset original, 28 responden a un perfil de riesgo alto. A continuación muestro las medidas más relevantes de las variables de estas 28 observaciones.

```
##        I              Edad             H                P
##  Min.   : 9.30   Min.   :19.00   Min.   :0.0000   Min.   :44.00
##  1st Qu.:13.85   1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:54.75
##  Median :14.65   Median :31.00   Median :1.0000   Median :57.00
##  Mean   :14.94   Mean   :30.04   Mean   :0.9643   Mean   :57.79
##  3rd Qu.:16.15   3rd Qu.:34.25   3rd Qu.:2.0000   3rd Qu.:61.00
##  Max.   :20.10   Max.   :52.00   Max.   :3.0000   Max.   :71.00
##        R         Sexo_HOMBRE Sexo_MUJER EC_SOLTERO EC_CASADO A_BAJO A_ALTO
##  Min.   :33.00   0:13        0:15       0:11       0:17      0:12   0:25
##  1st Qu.:42.00   1:15        1:13       1:17       1:11      1:16   1: 3
##  Median :49.50
##  Mean   :51.25
##  3rd Qu.:60.25
##  Max.   :68.00
##  A_MEDIO
##  0:19
##  1: 9
##
##
##
##
```

De acuerdo a la información de la tabla, los clientes que presentan un perfil de riesgo alto son generalmente de mediana edad, unos 30 años; con unos ingresos que oscilan entre 14.000 y 16.000 euros anuales por término medio. En su mayoría están solteros y aquellos que tienen hijos, suele ser sólo uno; presentan un ratio de endeudamiento sobre el patrimonio de aproximadamente el 50% y una baja aversión al riesgo.

A continuación compruebo la distribución de las variables numéricas asociadas a cada una de las clases de perfil de riesgo.



De los gráficos no se extrae de forma definida si las variables siguen una distribución normal, por ello procedo a realizar el test de normalidad de Shapiro-Wilk.

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Alto$I
## W = 0.96626, p-value = 0.4846

##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Alto$Edad
## W = 0.9052, p-value = 0.01518

##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Alto$H
```

```
## W = 0.82273, p-value = 0.000276

##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Alto$P
## W = 0.95356, p-value = 0.2432

##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Alto$R
## W = 0.94676, p-value = 0.1641
```

Para un nivel de significación del 5%, se comprueba que a excepción de la edad y el número de hijos, el resto de variables siguen una distribución normal.

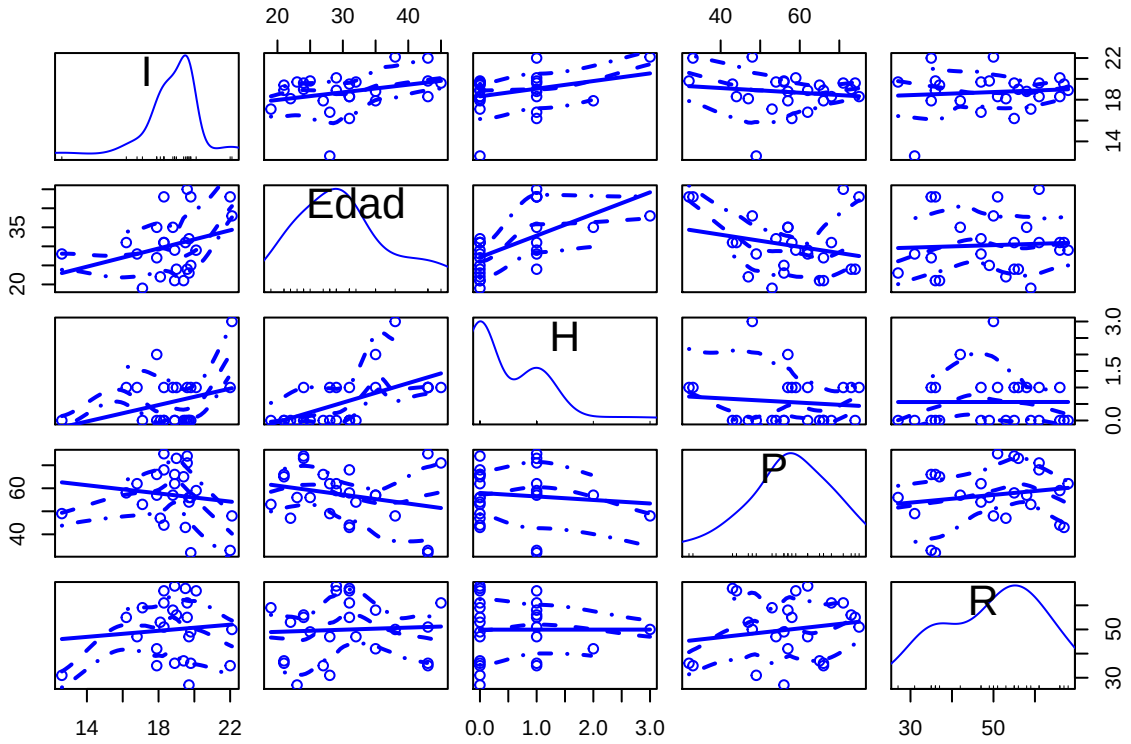
RIESGO MEDIO

Del conjunto de las 80 observaciones correspondientes al dataset original, 27 responden a un perfil de riesgo medio. A continuación muestro las medidas más relevantes de las variables de estas 27 observaciones.

```
##        I              Edad             H              P
##  Min.   :12.60   Min.   :19.00   Min.   :0.00   Min.   :32.00
##  1st Qu.:18.10   1st Qu.:24.00   1st Qu.:0.00   1st Qu.:49.00
##  Median :18.90   Median :29.00   Median :0.00   Median :57.00
##  Mean   :18.74   Mean   :30.32   Mean   :0.56   Mean   :57.12
##  3rd Qu.:19.70   3rd Qu.:35.00   3rd Qu.:1.00   3rd Qu.:66.00
##  Max.   :22.10   Max.   :45.00   Max.   :3.00   Max.   :75.00
##        R         Sexo_HOMBRE Sexo_MUJER EC_SOLTERO EC_CASADO A_BAJO A_ALTO
##  Min.   :27.00   0:15        0:10       0:14       0:11      0:19   0:22
##  1st Qu.:37.00   1:10        1:15       1:11       1:14      1: 6   1: 3
##  Median :51.00
##  Mean   :49.92
##  3rd Qu.:59.00
##  Max.   :68.00
##  A_MEDIO
##  0: 9
##  1:16
##
##
##
##
```

De acuerdo a la información de la tabla, los clientes que presentan un perfil de riesgo medio también son de mediana edad, unos 30 años; con unos ingresos que oscilan entre 18.000 y 20.000 euros anuales por término medio. Algo más de la mitad están casados y no suelen tener hijos; presentan un ratio de endeudamiento sobre el patrimonio de alrededor del 50% de este y una aversión al riesgo media.

Al igual que con el perfil de riesgo alto, con el medio compruebo la distribución de las variables.



Para asegurarme de si siguen una distribución normal las variables, llevo a cabo el test de Shapiro-Wilk.

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Medio$I
## W = 0.88465, p-value = 0.008632
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Medio$Edad
## W = 0.9346, p-value = 0.111
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Medio$H
## W = 0.70997, p-value = 9.955e-06
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Medio$P
## W = 0.96266, p-value = 0.47
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Medio$R
## W = 0.94683, p-value = 0.2125
```

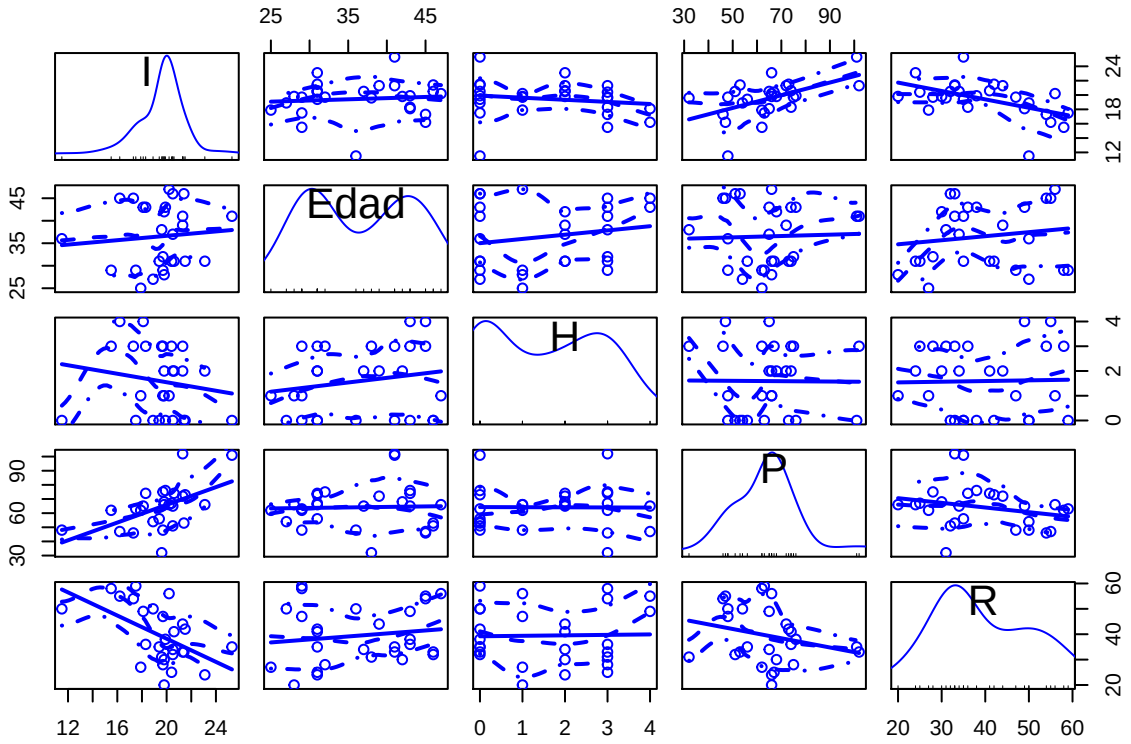
A excepción el número de hijos y los ingresos anuales, el resto de variables siguen una distribución normal de acuerdo al test de Shapiro-Wilk para un nivel de significatividad del 5%.

RIESGO BAJO

Del conjunto de las 80 observaciones correspondientes al dataset original, 25 responden a un perfil de riesgo bajo. A continuación muestro las medidas más relevantes de las variables de estas 25 observaciones.

```
##        I              Edad             H               P
##  Min.   :11.50   Min.   :25.00   Min.   :0.000   Min.   : 32.00
##  1st Qu.:18.20   1st Qu.:31.00   1st Qu.:0.000   1st Qu.: 53.50
##  Median :19.80   Median :37.00   Median :2.000   Median : 65.00
##  Mean   :19.45   Mean   :36.52   Mean   :1.593   Mean   : 64.15
##  3rd Qu.:20.55   3rd Qu.:43.00   3rd Qu.:3.000   3rd Qu.: 72.50
##  Max.   :25.30   Max.   :47.00   Max.   :4.000   Max.   :102.00
##        R         Sexo_HOMBRE Sexo_MUJER EC_SOLTERO EC_CASADO A_BAJO A_ALTO
##  Min.   :20.00   0:13        0:14       0:16       0:11      0:18   0:18
##  1st Qu.:31.50   1:14        1:13       1:11       1:16      1: 9   1: 9
##  Median :36.00
##  Mean   :39.48
##  3rd Qu.:49.50
##  Max.   :59.00
##  A_MEDIO
##  0:18
##  1: 9
##
##
##
##
```

De acuerdo a la información de la tabla, los clientes que presentan un perfil de riesgo bajo presentan una edad que oscila entre los 30 y 40 años; con unos ingresos que van de 18.000 a 20.500 anuales por término medio. Algo más de la mitad están casados y tienen alrededor de 2 hijos; presentan un ratio de endeudamiento sobre el patrimonio del 65% aproximadamente y la aversión al riesgo se encuentra dividida a partes iguales entre baja, media y alta.



Llevo a cabo la prueba de Shapiro-Wilk para comprobar la normalidad.

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Bajo$I
## W = 0.92551, p-value = 0.05369
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Bajo$Edad
## W = 0.90995, p-value = 0.02276
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Bajo$H
## W = 0.86166, p-value = 0.001977
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Bajo$P
## W = 0.9327, p-value = 0.0805
```

```
##
##  Shapiro-Wilk normality test
##
## data:  riesgo_Bajo$R
## W = 0.95138, p-value = 0.2316
```

Para el perfil de riesgo bajo se comprueba que la edad y el número de hijos no siguen una distribución normal de acuerdo al test de Shapiro-Wilk, con un nivel de significatividad del 5%.

De acuerdo a los test realizados existen evidencias de falta de normalidad para algunas de las variables de cada grupo de riesgo. Supongo que la matriz de covarianzas es igual para todos los grupos.

Realizo el análisis discriminante, aún así, ya sé que la primera de las condiciones no se cumple, por tanto no es plenamente adecuado en este caso.

ANALISIS DISCRIMINANTE

Para la realización del análisis discriminante tendré únicamente en cuenta las variables numéricas del dataset. Es decir, bajo mi punto de vista, ni el estado civil, ni el sexo, ni el grado de aversión al riesgo tienen influencia sobre el perfil de riesgo de los clientes. Si lo tienen sin embargo los ingresos anuales, el patrimonio, el número de hijos a su cargo, la edad y el nivel de endeudamiento. La edad de una persona determina la experiencia profesional de esta, lo que puede influir en el nivel de ingresos anuales y estos a su vez en el patrimonio que cada persona posee. El número de hijos, a mi parecer influye en el nivel de gasto existente de cada cliente, el cual puede influir en el perfil de riesgo.

LDA

El objetivo del LDA es generar combinaciones lineales de las variables originales que ofrezcan la mejor separación posible entre los 3 grupos de riesgo que tenemos en nuestro dataset. Debo considerar los 3 grupos que hay y las 5 variables numéricas, el número máximo de funciones discriminantes válidas será el mínimo entre el número de grupos menos uno y el número de variables.

A continuación visualizo los valores de las cargas de las funciones discriminantes.

```
## Call:
## lda(TIPO ~ ., data = datos)
##
## Prior probabilities of groups:
##  Alto riesgo  Bajo riesgo Riesgo medio
##       0.3500       0.3375       0.3125
##
## Group means:
##                       I     Edad         H        P        R
## Alto riesgo  14.93929 30.03571 0.9642857 57.78571 51.25000
## Bajo riesgo  19.45185 36.51852 1.5925926 64.14815 39.48148
## Riesgo medio 18.73600 30.32000 0.5600000 57.12000 49.92000
##
## Coefficients of linear discriminants:
##               LD1         LD2
## I     0.337106536  0.27379153
## Edad  0.026494684 -0.05813062
## H     0.220649027 -0.37834862
## P     0.004647734 -0.03571364
```

```
## R    -0.034172832  0.04067145
##
## Proportion of trace:
##   LD1   LD2
## 0.744 0.256
```

Cada una de las funciones discriminantes es una combinación lineal de las variables ingresos anuales, edad, número de hijos, patrimonio y ratio de endeudamiento del patrimonio. Así pues, la primera función discriminante queda definida de la siguiente forma:

```
0.337*I + 0.026*Edad + 0.220*H +0.004*P - 0.034*R
```

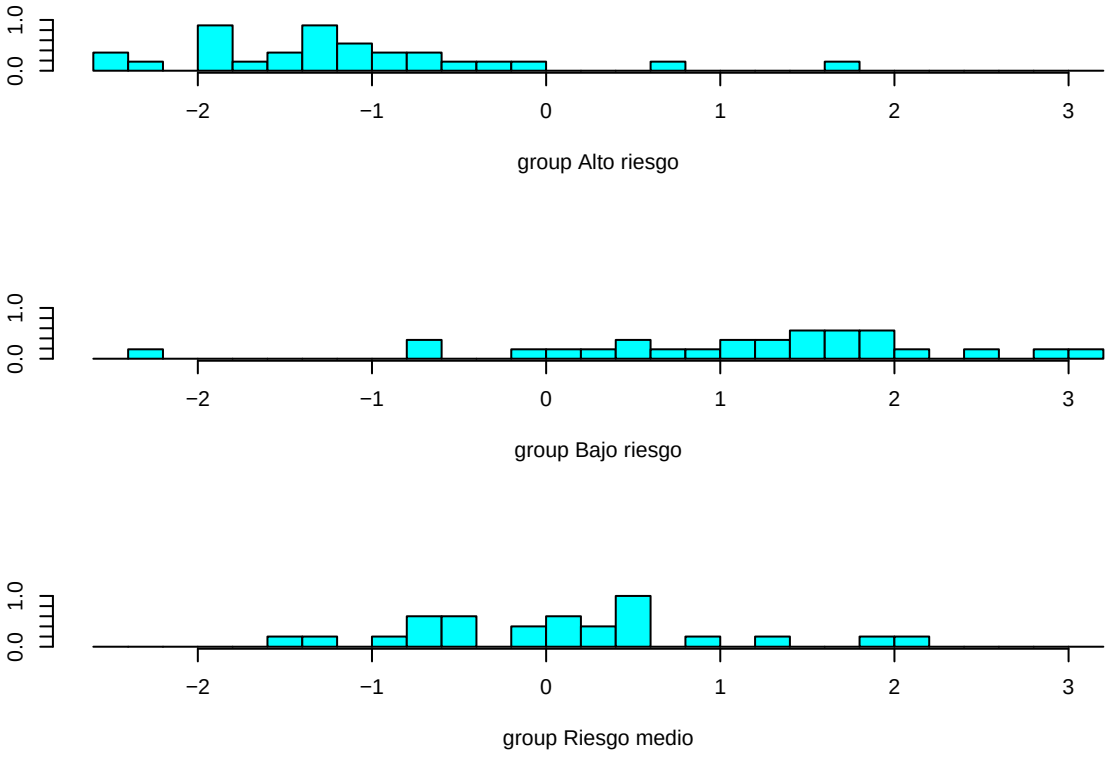
Mientras que la segunda función discriminante es:

```
0.273*I - 0.058*Edad - 0.378*H - 0.035*P + 0.040*R
```

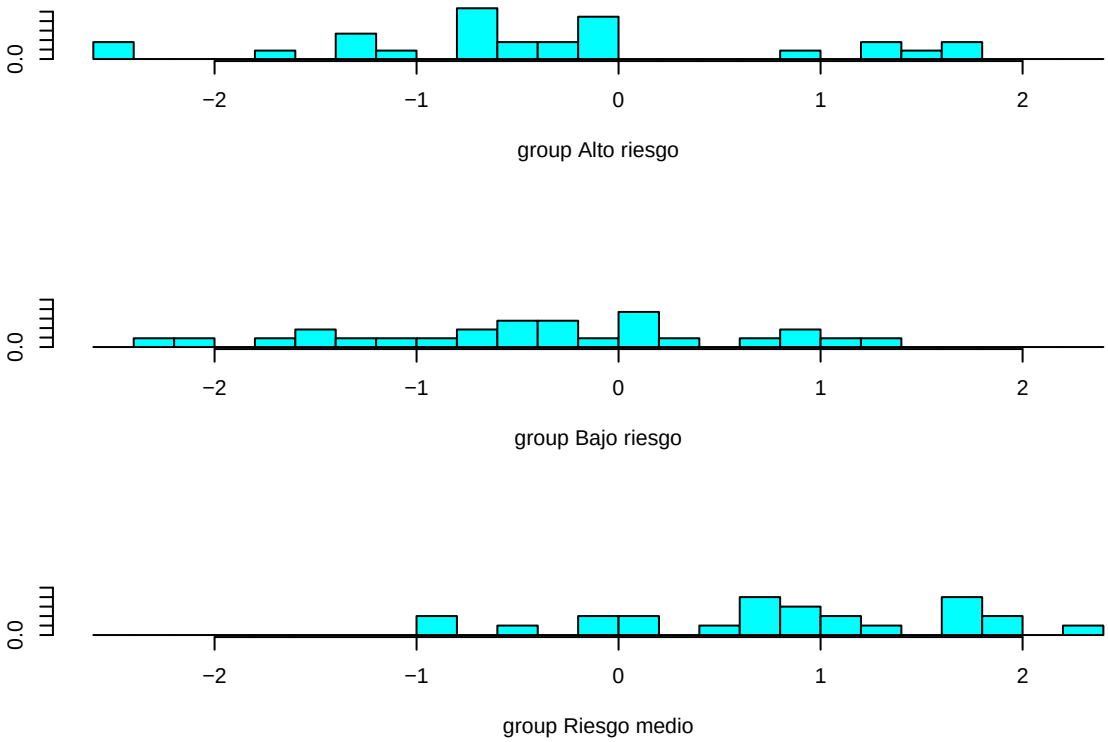
La primera función discriminante consigue un tanto por ciento de separación del 74,4%. Mientras que la segunda función discriminante únicamente consigue un 25,6% de separación.

Ahora calculo un vector de predicción con 2 dimensiones, una para cada grupo. Procedo a representar gráficamente cada una de las dos funciones discriminantes calculadas, de esta forma observo como trabaja cada una y si realiza una correcta diferenciación entre los 3 grupos de riesgo, bajo, medio y alto.

PRIMERA FUNCION DISCRIMINANTE



SEGUNDA FUNCION DISCRIMINANTE



La primera función discriminante diferencia bien entre el grupo de riesgo alto y el bajo, sin embargo con el grupo de riesgo medio comete errores. La segunda función discriminante comete errores entre los 3 grupos.

MATRIZ DE CONFUSION Y PREDICCION

Ahora dibujo la matriz de confusión y hallo la precisión del modelo.

```
##
##                Alto riesgo Bajo riesgo Riesgo medio
##   Alto riesgo           21           2            3
##   Bajo riesgo            2          20            6
##   Riesgo medio           5           5           16
```

Como medida de la precisión del modelo calculo el error cometido por este en la predicción de los grupos de riesgo.

El modelo comete un error de un 28,75% en la predicción del grupo de riesgo de los clientes.

Realizo un análisis discrimante cuadrático.

QDA

Procedo a la realización del análisis discriminante cuadrático.

```
## Call:
## qda(TIPO ~ ., data = datos)
##
## Prior probabilities of groups:
##  Alto riesgo  Bajo riesgo Riesgo medio
##       0.3500       0.3375       0.3125
##
## Group means:
##                       I     Edad         H        P        R
## Alto riesgo  14.93929 30.03571 0.9642857 57.78571 51.25000
## Bajo riesgo  19.45185 36.51852 1.5925926 64.14815 39.48148
## Riesgo medio 18.73600 30.32000 0.5600000 57.12000 49.92000
```

La matriz de confusión es la siguiente:

```
##
##                Alto riesgo Bajo riesgo Riesgo medio
##   Alto riesgo           22           2            4
##   Bajo riesgo            2          20            3
##   Riesgo medio           4           5           18
```

Hallo el riesgo cometido por el modelo calculado.

El error cometido por el modelo mediante el análisis discriminante cuadrático es de un 25%. Se obtiene una mayor precisión en la predicción de los grupos de riesgo de los clientes mediante el QDA.

ARBOL DE DECISION

Lo primero ha realizar es una división de la muestra en train y validation, esta se realiza con una proporción 80:20. El criterio de decisión en la elección de las variables de cara a calcular el árbol de clasificación es el mismo que he aplicado anteriormente para el análisis discriminante.

Compruebo que la división del conjunto de observaciones se realiza de forma balanceada. Muestro las observaciones del conjunto train en primer lugar y posteriormente del validation.

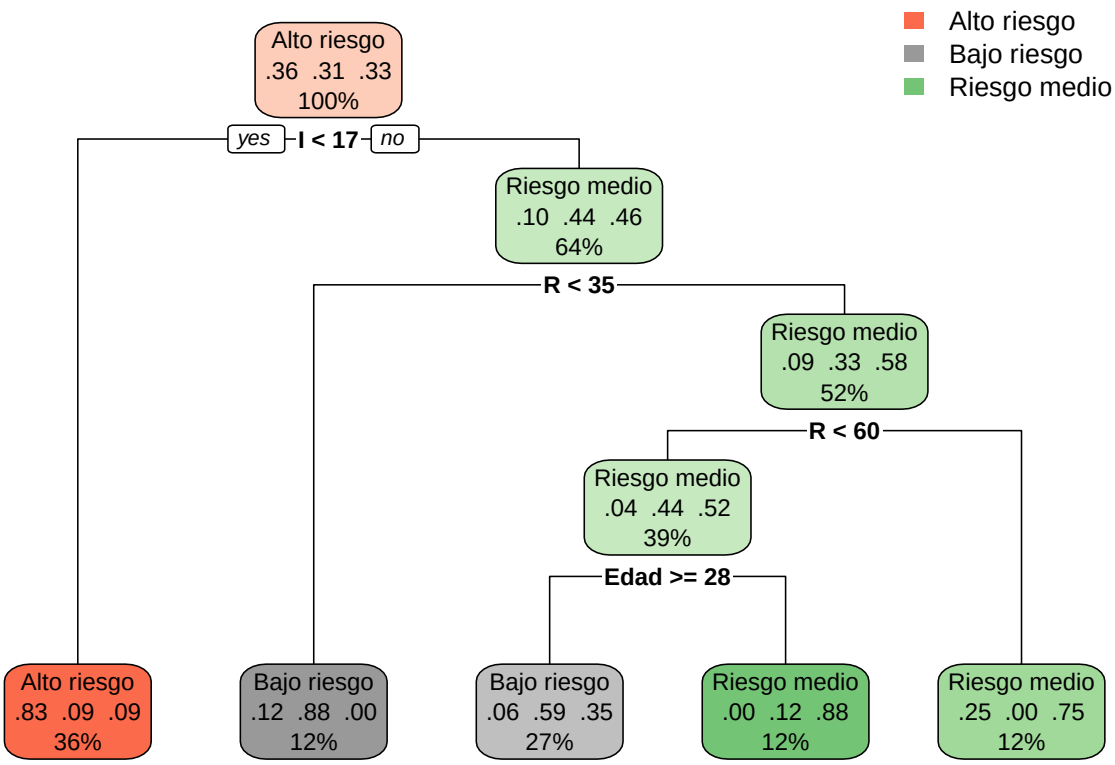
```
##
##  Alto riesgo  Bajo riesgo Riesgo medio
##           23           20           21
```

```
##
##  Alto riesgo  Bajo riesgo Riesgo medio
##            5            7            4
```

La muestra como se puede comprobar está balanceada, por tanto, procedo a la estimación del árbol de inferencia.

ESTIMACION, REPRESENTACION E INTERPRETACION

Para poder interpretar el árbol y observar los nodos, represento este gráficamente.



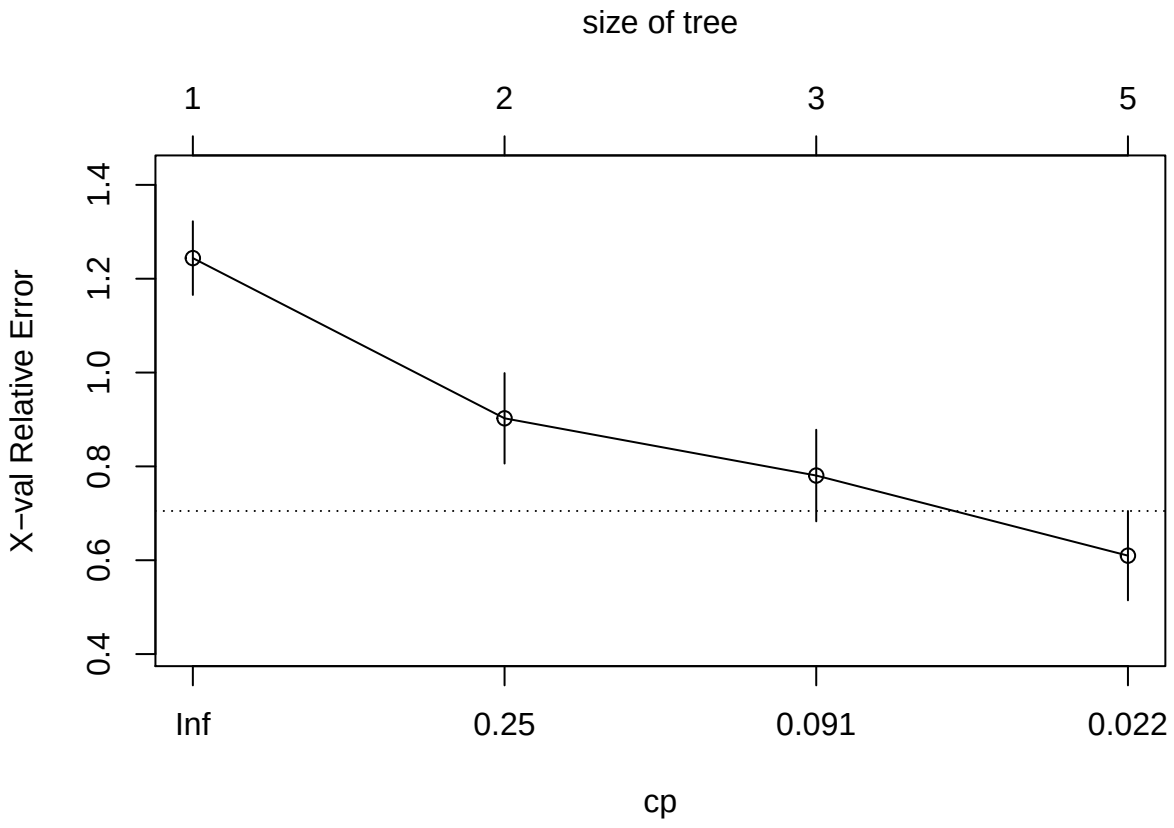
Las variables de mayor importancia son los ingresos, el ratio de endeudamiento sobre el patrimonio y la edad, estas son las utilizadas en el árbol. En primer lugar se evalúa el nivel de ingresos anuales, si es menor a 17.000 euros anuales directamente se asigna al cliente como de alto riesgo con un porcentaje de precisión del 83%. En caso de tener unos ingresos superiores a 17.000 euros anuales, en principio se le asigna a un perfil de riesgo medio, sin embargo, el nodo correspondiente no diferencia claramente entre bajo riesgo y medio, por ello se utiliza entonces el ratio de endeudamiento. Si el ratio de endeudamiento es inferior al 35% se asigna al cliente como de bajo riesgo, con una precisión del 88%. En caso contrario, se le asigna como de riesgo medio, con una precisión del 58%. Se vuelve a evaluar el ratio de endeudamiento, si este es inferior al 60% se le asigna a un perfil de riesgo medio con una precisión del 52%, en este nodo entonces se evalúa la edad. Si es inferior a los 28 años se le asigna como de bajo riesgo con una precisión del 59%, en caso contrario es un perfil de riesgo medio con un 88% de precisión. En el caso de que los ingresos sean superiores a 17.000 euros anuales y el ratio de endeudamiento sea superior al 60%, se establece al cliente como de riesgo medio con una precisión del 75%.

El siguiente paso será llevar a cabo el cálculo del árbol podado, para ello obtengo el grado de complejidad paramétrica. Debo de establecer el grado de complejidad paramétrica que presente un menor error, aunque he de atender a que el error +/- la desviación típica no de como resultado un error total superior al de un grado de complejidad paramétrica inferior.

Obtengo la tabla de complejidad paramétrica.

```
##           CP nsplit rel error    xerror       xstd
## 1 0.36585366      0 1.0000000 1.2439024 0.07850240
## 2 0.17073171      1 0.6341463 0.9024390 0.09636269
## 3 0.04878049      2 0.4634146 0.7804878 0.09756098
## 4 0.01000000      4 0.3658537 0.6097561 0.09519814
```

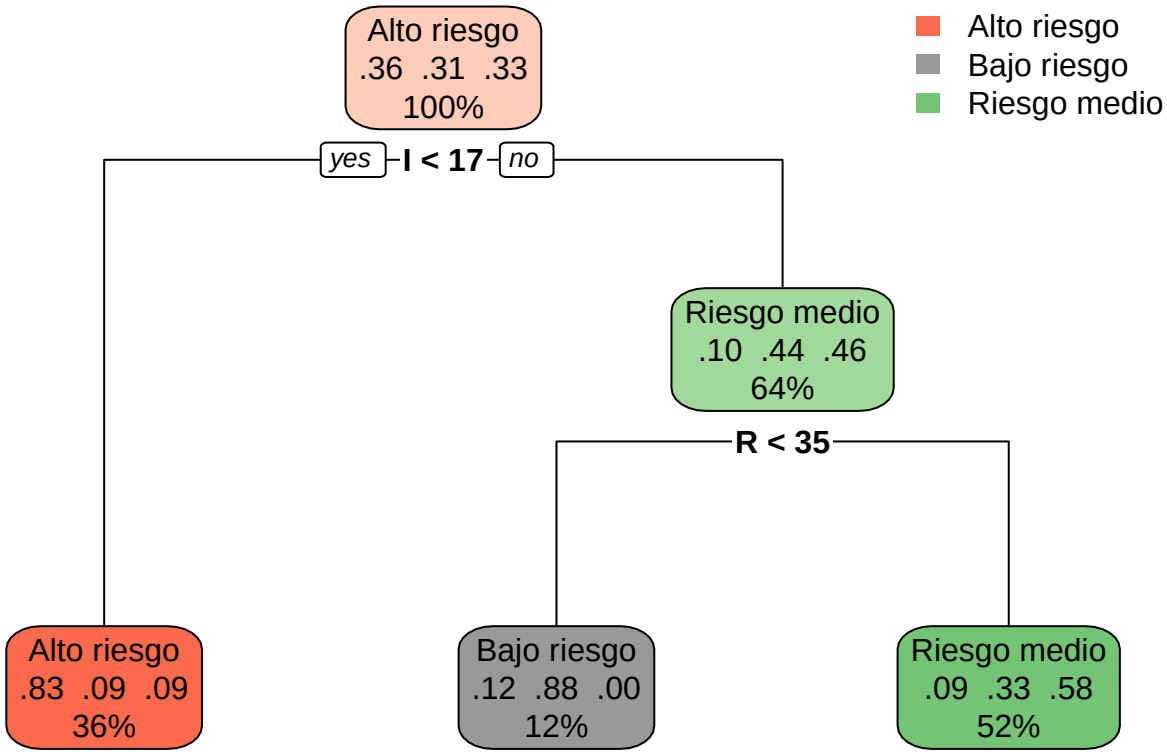
Para visualizar mejor que grado de complejidad paramétrica escoger, realizo un gráfico de la curva.



Dados los resultados obtenidos mediante la tabla de complejidad paramétrica y la visualización de la curva, determino que he de quedarme con 4 grados de complejidad. Por tanto, no tendría sentido realizar una poda del árbol, aún así llevaré a cabo una poda y estableceré 3 grados de complejidad paramétrica con el objetivo de realizar una comparación entre los árboles una vez hechas las predicciones sobre la muestra de validación.

La complejidad paramétrica para la poda del árbol es de 0.04878049.

Represento el árbol podado.



El nuevo árbol podado únicamente tiene en cuenta los ingresos anuales y el ratio de endeudamiento. Si los ingresos son inferiores a 17.000 euros anuales, el cliente queda establecido dentro del grupo de alto riesgo con una precisión del 83%. En caso contrario se procede a valuar el ratio de endeudamiento. Si este es inferior al 35% se le asigna al grupo de bajo riesgo con un precisión del 88%, si es superior al 35% se le engloba dentro del grupo de riesgo medio con una precisión del 58%.

El siguiente pasó será realizar la predicción tanto del árbol original, como del podado, sobre la muestra de validación, representar las matrices de confusión y establecer el porcentaje de error de cada árbol.

ARBOL SIN PODAR

La matriz de confusión del árbol original, es decir, sin podar; será la siguiente:

```
##                Predicted
## Actual          Alto riesgo Bajo riesgo Riesgo medio
##   Alto riesgo             4           0            1
##   Bajo riesgo             1           6            0
##   Riesgo medio            1           2            1
```

El error cometido por el árbol original es de un 31,25%.

ARBOL PODADO

La matriz de confusión del árbol podado es la siguiente.

```
##               Predicted
## Actual         Alto riesgo Bajo riesgo Riesgo medio
##   Alto riesgo            4           0            1
##   Bajo riesgo            1           4            2
##   Riesgo medio           1           1            2
```

El error cometido por el árbol podado es de un 37,5%.

La principal cuestión a la hora de realizar el podado de los árboles es ver qué interesa más, perder precisión y ganar simplicidad, o lo contrario.

En este caso, a mi parecer, la pérdida de precisión no compensa la ganancia de simplicidad del árbol.

REFERENCIAS

- Código en GitHub: https://github.com/BeltranCunef/Master