EXAMEN CLASIFICACIÓN

JORGE CASAN VÁZQUEZ

5 de febrero de 2019

Objetivo: El presente informe emplea los datos de la Encuesta de Presupuestos Familiares 2017 y corresponde a hogares single (solteros),con la finalidad de explicar el consumo cárnico a partir de la determinación de las variables explicativas; en este sentido se propone la estimación de un modelo explicativo por medio de dos modelos. En primer lugar, la regresión logística; en segundo lugar, a través de los árboles de clasificación, para finalmente realizar una comparación entre ellos estableciendo una jerarquía según la cual veremos qué modelo es el que realiza una mejor clasificación. Dicha comparación la realizaré calculando para los 2 modelos las matrices de confusión respectivas, la precisión del modelo y el área bajo la curva calculada mediante el análisis ROC, tanto para la variable CAT2 como CAT3.

La encuesta de Presupuestos familiares (EPF) la realiza cada año en Instituto Nacional de Estadística (INE),siendo su objetivo en disponer de estadísticas comparativas con el ánimo de conocer el gasto en consumo de los hogares residentes en España, así como la distribución del mismo entre las diferentes parcelas de consumo, sustituyendo a la Encuesta Contínua de Presupuestos Familiares (ECPF) que estuvo en vigor desde el año 1997 al 2005 incorporando diversas mejoras metodológicas , tales como el cambio de periodicidad (de trimestral a anual), así como el aumento del tamaño de la muestra.

La EPF ofrece la información imprescindible para las estimaciones sobre el gasto en consumo de los hogares de la Contabilidad Nacional y para la actualización de ponderaciones del Índice de Precios al Consumo (IPC).

Los gastos de consumo que se registran en la EPF se refieren tanto al flujo monetario que destina el hogar, en este caso, single, al pago de determinados bienes y servicios de consumo final asi como al valor de determinados consumos no monetarios efectuados por los hogares. Estos últimos son por ejemplo nuestra variable 'REGTEN'la cual tomará valor *0* en caso de alquiler imputado/pago de hipoteca.

Las variables objeto de estudio son las siguientes:

Table 1: Tabla de las variables objeto de estudio

VARIABLES	DEFINICIÓN
cat2	Variable de clasificación de hogares según su gasto en consumo de vacuno anual
cat3	Variable de clasificación de hogares según su gasto en consumo de vacuno anual
TAMAMU	Tamaño de los municipios
DENSIDAD	Densidad de la población
EDAD	Edad, expresada en años
SEXO	Sexo de la muestra
ESTUD	Nivel de estudios completados
LAB	Situación laboral
REGTEN	Regimen de tenencia de la vivienda
SUPERF	Superficie de la vivienda en metros cuadrados
IMPEXAC	Importe exacto de los ingresos mensuales netos totales del hogar en cientos de €

Cabe destacar que la obtención de los datos se realiza a partir del muestreo de la población, en donde se aplica una muestra de 4220 hogares solteros distribuidas por todo el territorio nacional.Para comenzar, se plantea el análisis exploratorio de los datos con la finalidad de identificar las variables explicativas relacionadas con la variable dependiente.

PREGUNTA 1: ANÁLISIS EXPLORATORIO

Realizamos el análisis eploratorio con el objetivo de identificar las variables explicativas con la variable dependiente.

De las 11 variables observamos que la variable SUPERF tiene 168 valores perdidos, representando casi un 4% sobre el total de las observaciones para esa variable. Lo mejor sería reemplazar todos los valores perdidos por su mediana, puesto que considero que es la forma más representativa de reemplzar los NA´s y no aumentemos la dispersión.

Nos creamos una semilla 1234, para que cada vez que se carguen los datos, estos no den diferentes resultados y clasificamos el conjunto de las 4220 observaciones en dos subconjuntos que serán por una parte, los datos_train los cuales albergarán el 80% del total de observaciones y datos_test el resto.

Pasaremos a factor las variables que nos interesen y dejaremos como están las variables SUPERF y EDAD, puesto que no tendría mucho sentido convertirlas a factor. Por otra parte, la variable EDAD la pasaremos como numérica.

```
library(readxl)

datos<- read_xlsx('BDexamen1.xlsx', sheet='bd', col_names = TRUE)
# Exploration variables
str(datos)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    4220 obs. of  11 variables:
##  $ TAMAMU  : num  0 1 1 0 1 1 1 0 1 1 ...
##  $ DENSIDAD: num  2 1 2 3 2 2 1 3 1 1 ...
##  $ EDAD    : chr  "74" "51" "54" "42" ...
##  $ SEXO    : num  0 1 0 1 0 0 0 0 1 0 ...
##  $ ESTUD   : num  2 2 1 1 3 2 2 2 4 1 ...
##  $ LAB     : num  4 3 3 1 3 3 3 4 1 4 ...
##  $ REGTEN  : num  1 1 1 1 1 1 0 0 0 0 ...
##  $ SUPERF  : num  78 78 91 150 90 90 60 75 95 50 ...
##  $ IMPEXAC : num  6.57 0 4.26 0 0 4.26 5.5 7.43 3.5 7.42 ...
##  $ cat2    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ cat3    : num  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(datos)
```

```
##      TAMAMU          DENSIDAD         EDAD                SEXO
##  Min.   :0.0000   Min.   :1.000   Length:4220        Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.000   Class :character   1st Qu.:0.0000
##  Median :1.0000   Median :1.000   Mode  :character   Median :0.0000
##  Mean   :0.7758   Mean   :1.746                      Mean   :0.4116
##  3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :3.000                      Max.   :1.0000
##
##      ESTUD            LAB            REGTEN           SUPERF
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   : 35.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 67.00
##  Median :2.000   Median :4.000   Median :1.0000   Median : 85.00
##  Mean   :2.437   Mean   :2.872   Mean   :0.6751   Mean   : 90.95
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:100.00
##  Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :300.00
##                                                   NA's   :168
##     IMPEXAC            cat2             cat3
##  Min.   :  0.00   Min.   :0.0000   Min.   :1.000
```

```
##  1st Qu.:  7.45   1st Qu.:0.0000   1st Qu.:1.000
##  Median : 10.05   Median :0.0000   Median :2.000
##  Mean   : 11.84   Mean   :0.4763   Mean   :1.953
##  3rd Qu.: 15.00   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :152.07   Max.   :1.0000   Max.   :3.000
##
```

```
# Explore NA values
ExploreNA <- function(datos) {
  TrueNA <- is.na.data.frame(datos)
  SumNA <- colSums(TrueNA)
  PorcentNA <- colSums(TrueNA) / nrow(datos)*100
  VariableNA <- data.frame(SumNA, PorcentNA)

  return(VariableNA)
}
ExploreNA(datos)
```

```
##          SumNA PorcentNA
## TAMAMU       0  0.000000
## DENSIDAD     0  0.000000
## EDAD         0  0.000000
## SEXO         0  0.000000
## ESTUD        0  0.000000
## LAB          0  0.000000
## REGTEN       0  0.000000
## SUPERF     168  3.981043
## IMPEXAC      0  0.000000
## cat2         0  0.000000
## cat3         0  0.000000
```

```
f=function(x){
  x<-as.numeric(as.character(x)) #first convert each column into numeric if it is from factor
  x[is.na(x)] =median(x, na.rm=TRUE) #convert the item with NA to median value from the column
  x #display the column
}
datos=data.frame(apply(datos,2,f))
summary(datos)
```

```
##      TAMAMU          DENSIDAD          EDAD            SEXO
##  Min.   :0.0000   Min.   :1.000   Min.   :18.00   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:50.00   1st Qu.:0.0000
##  Median :1.0000   Median :1.000   Median :63.00   Median :0.0000
##  Mean   :0.7758   Mean   :1.746   Mean   :61.57   Mean   :0.4116
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:75.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :3.000   Max.   :85.00   Max.   :1.0000
##      ESTUD            LAB           REGTEN          SUPERF
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   : 35.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 68.75
##  Median :2.000   Median :4.000   Median :1.0000   Median : 85.00
##  Mean   :2.437   Mean   :2.872   Mean   :0.6751   Mean   : 90.72
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:100.00
##  Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :300.00
##     IMPEXAC            cat2            cat3
##  Min.   :  0.00   Min.   :0.0000   Min.   :1.000
```

```
##  1st Qu.:  7.45   1st Qu.:0.0000   1st Qu.:1.000
##  Median : 10.05   Median :0.0000   Median :2.000
##  Mean   : 11.84   Mean   :0.4763   Mean   :1.953
##  3rd Qu.: 15.00   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :152.07   Max.   :1.0000   Max.   :3.000
```

```
# Create dummy variables (function)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)

#Conversión a factor de las variables que nos interesen

#Para nuestra variable dependiente, la cual será CAT2 y CAT3
datos$cat2 <- as.factor(datos$cat2)
datos$cat3 <- as.factor(datos$cat3)

#Para el resto de las variables
datos$TAMAMU <- as.factor(datos$TAMAMU)
datos$DENSIDAD <- as.factor(datos$DENSIDAD)
datos$SEXO <- as.factor(datos$SEXO)
datos$ESTUD <- as.factor(datos$ESTUD)
datos$LAB <- as.factor(datos$LAB)
datos$REGTEN <- as.factor(datos$REGTEN)
datos$EDAD<- as.numeric(datos$EDAD)

#No factorizarmos las variables EDAD, SUPERF, y IMPEXAC

str(datos)
```

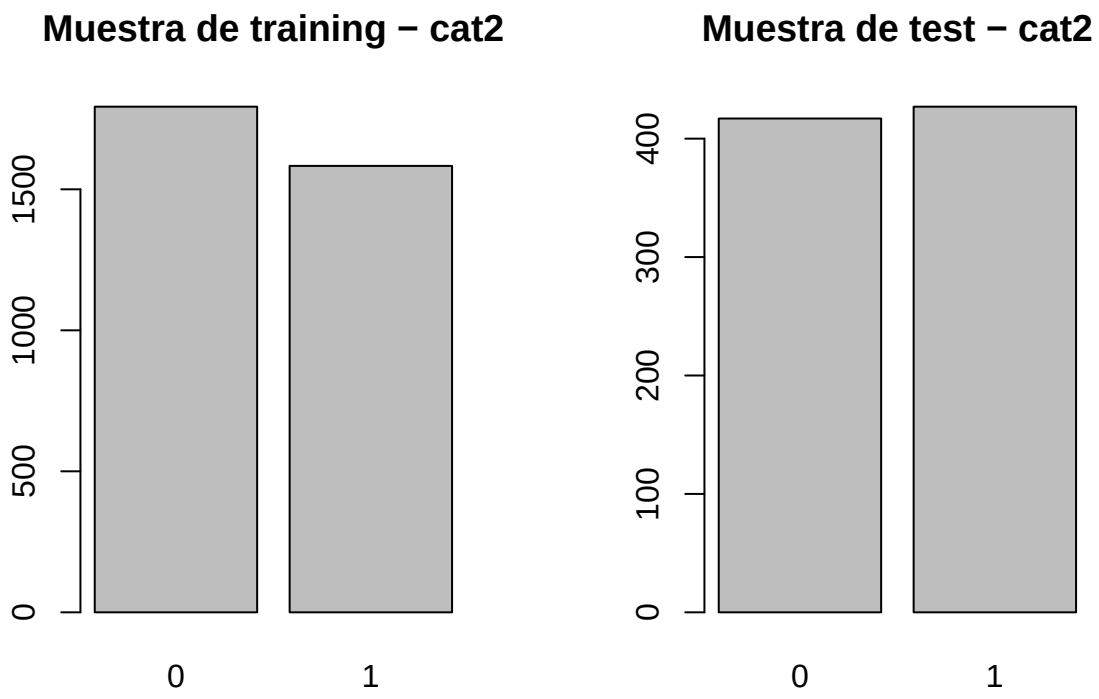
```
## 'data.frame':    4220 obs. of  11 variables:
##  $ TAMAMU  : Factor w/ 2 levels "0","1": 1 2 2 1 2 2 2 1 2 2 ...
##  $ DENSIDAD: Factor w/ 3 levels "1","2","3": 2 1 2 3 2 2 1 3 1 1 ...
##  $ EDAD    : num  74 51 54 42 54 58 45 72 41 85 ...
##  $ SEXO    : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 1 1 2 1 ...
##  $ ESTUD   : Factor w/ 4 levels "1","2","3","4": 2 2 1 1 3 2 2 2 4 1 ...
##  $ LAB     : Factor w/ 4 levels "1","2","3","4": 4 3 3 1 3 3 3 4 1 4 ...
##  $ REGTEN  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 1 1 1 ...
##  $ SUPERF  : num  78 78 91 150 90 90 60 75 95 50 ...
##  $ IMPEXAC : num  6.57 0 4.26 0 0 4.26 5.5 7.43 3.5 7.42 ...
##  $ cat2    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cat3    : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Divide train and test sample
set.seed(1234)
train <- sample(nrow(datos), 0.8*nrow(datos))
```

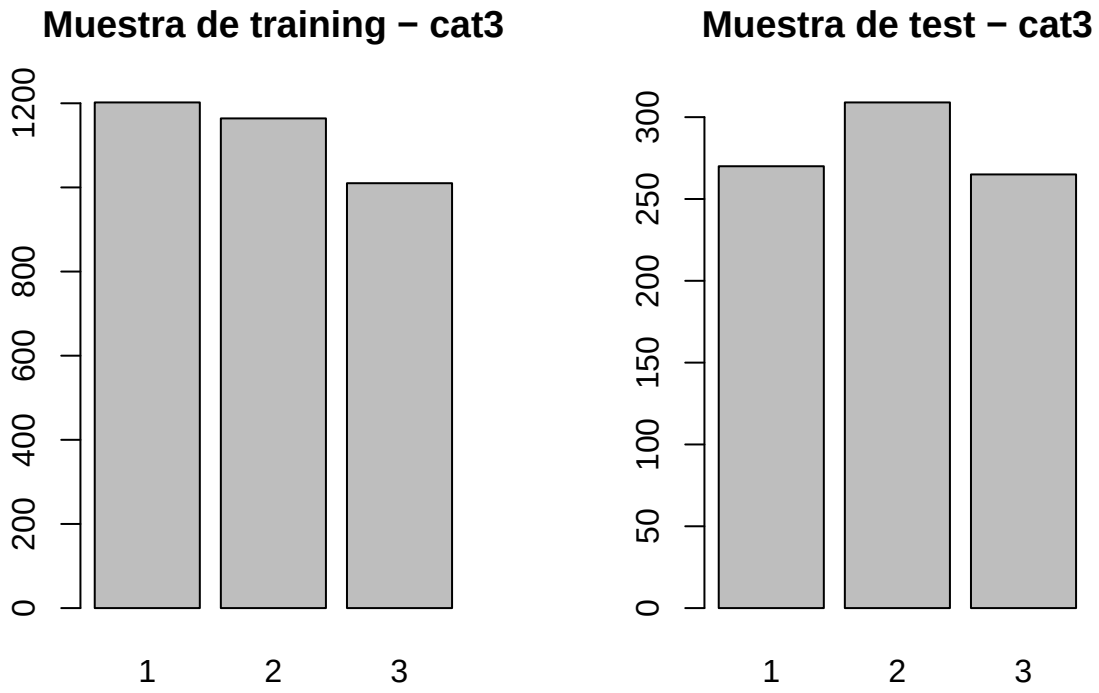
```
datos_train <- datos[train,]
datos_test <- datos[-train,]
```

Vamos a ver cómo están de balanceados nuestros datos para la variable dependiente, tanto en la muestra con que trabajaremos denominada datos_train como en el test. Observamos que ambas están balanceadas, existiendo una proporcion superior de hogares cuyo consumo de carne vacuna anual es baja, tanto para el train como para el test y para ambas variables, esto es, para la variable cat2 y cat3.

```
par(mfrow = c(1,2))
plot(as.factor(datos_train$cat2), main = "Muestra de training - cat2")
plot(as.factor(datos_test$cat2), main = "Muestra de test - cat2")
```




```
par(mfrow = c(1,2))
plot(as.factor(datos_train$cat3), main = "Muestra de training - cat3")
plot(as.factor(datos_test$cat3), main = "Muestra de test - cat3")
```



En este punto, ya está todo preparado para proceder a realizar una regresión logística y un árbol de clasificación para predecir las variables dependientes.

PREGUNTA 2: Aplicar un modelo de regresión logística con CAT2 y un árbol de clasificación

REGRESIÓN LOGÍSTICA

En esta parte del desarrollo del problema se va a realizar una regresión logística. Esta regresión re realizará para clasificar la variable cat2. Por ello, se elimina del dataset la variable cat3, pues no tendría sentido realizar una regresión logística con esa variable, para prevenir el sobreajuste (overfitting) y por otra parte evitando la multicolinealidad entre las variables.

Dado la variable que define el consumo de carne vacuna anual Cat2 es una variable dicotómica, indicando 0 si el consumo es bajo y 1 por otra parte si el consumo no es bajo, realizaremos la regresión logística empleando los datos de entrenamiento.

```
datos_train_1 <- datos_train[,-11]
datos_test_1 <- datos_test[,-11]
regresion <- glm(cat2 ~ ., family = "binomial", data = datos_train_1)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(regresion)
```

```
##
## Call:
## glm(formula = cat2 ~ ., family = "binomial", data = datos_train_1)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0573  -0.4206  -0.0207   0.5279   3.5450
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.433651   0.543395 -15.520  < 2e-16 ***
## TAMAMU1     -0.143995   0.173946  -0.828  0.40778
## DENSIDAD2   -0.583302   0.143340  -4.069 4.71e-05 ***
## DENSIDAD3   -0.472925   0.181121  -2.611  0.00903 **
## EDAD         0.003601   0.006114   0.589  0.55583
## SEXO1       -0.311327   0.115705  -2.691  0.00713 **
## ESTUD2       0.214781   0.135045   1.590  0.11174
## ESTUD3       0.171401   0.200806   0.854  0.39335
## ESTUD4       0.065569   0.219277   0.299  0.76492
## LAB2        -1.602606   0.364300  -4.399 1.09e-05 ***
## LAB3        -2.174590   0.287451  -7.565 3.88e-14 ***
## LAB4        -1.333037   0.239003  -5.577 2.44e-08 ***
## REGTEN1      7.115715   0.309986  22.955  < 2e-16 ***
## SUPERF       0.002598   0.001423   1.826  0.06788 .
## IMPEXAC      0.342325   0.018152  18.859  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4667.1  on 3375  degrees of freedom
## Residual deviance: 2291.7  on 3361  degrees of freedom
## AIC: 2321.7
##
## Number of Fisher Scoring iterations: 7
```

Con este modelo de regresión obtenemos un AIC de 2321.7, a través de la regresion logística empleando el método *both* en donde se combinan el método forward y backward obtendremos el modelo de regresión con las variables estádisticamente más significativas y el modelo que menor AIC tiene, siendo por ello el modelo de regresión logística que mejor ajuste tiene.

Obtenemos con esta técnica el menor AIC, siendo el más bajo de 2315. Realizamos el modelo de regresion con las variables estadísticamente más significativas.

```
regresion_buena <- glm(cat2 ~ DENSIDAD + SEXO+ LAB + REGTEN + SUPERF+  IMPEXAC, family ="binomial", data
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(regresion_buena)
```

```
##
## Call:
## glm(formula = cat2 ~ DENSIDAD + SEXO + LAB + REGTEN + SUPERF +
##     IMPEXAC, family = "binomial", data = datos_train_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0863  -0.4238  -0.0204   0.5353   3.5336
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.351405   0.410109 -20.364  < 2e-16 ***
## DENSIDAD2   -0.563593   0.140199  -4.020 5.82e-05 ***
## DENSIDAD3   -0.382296   0.127058  -3.009  0.00262 **
## SEXO1       -0.310372   0.113882  -2.725  0.00642 **
## LAB2        -1.593821   0.363758  -4.382 1.18e-05 ***
## LAB3        -2.163732   0.286818  -7.544 4.56e-14 ***
## LAB4        -1.309595   0.192064  -6.819 9.20e-12 ***
## REGTEN1      7.189994   0.307664  23.370  < 2e-16 ***
## SUPERF       0.002753   0.001413   1.948  0.05136 .
## IMPEXAC      0.343742   0.017590  19.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4667.1  on 3375  degrees of freedom
## Residual deviance: 2295.3  on 3366  degrees of freedom
## AIC: 2315.3
##
## Number of Fisher Scoring iterations: 7
```

Con un AIC de 2315.3 vemos que las variables estádisticamente más significativas son la densidad de la población, el sexo, la situación laboral del individuo, el régimen de tenencia,la superficie y los ingresos netos mensuales.

Con el análisis de los párametros de las variables, vemos que la densidad de la población DENSIDAD2 (zona intermedia) y DENSIDAD3 (zona diseminada), así como SEXO1 (hombre) son aquellas que mejor explican la variable predictora, por lo que guarda una estrecha relación entre ambas. Un aumento de esta variable explicativa contribuye a un aumento del consumo de la carne vacuna.

```
exp(coef(regresion_buena))
```

```
##  (Intercept)    DENSIDAD2    DENSIDAD3        SEXO1         LAB2
## 2.360646e-04 5.691602e-01 6.822933e-01 7.331740e-01 2.031479e-01
##         LAB3         LAB4      REGTEN1       SUPERF      IMPEXAC
## 1.148955e-01 2.699294e-01 1.326095e+03 1.002757e+00 1.410215e+00
```

El modelo presenta un R2 McFadden de 0.5081882 la cual debe tender a 1; por lo cual este valor se considera aceptable. Cuanto mas se acerque a 1 más ajustado será el modelo.

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
pR2(regresion_buena)
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -1147.6572428 -2333.5292649  2371.7440443     0.5081882     0.5046699
##          r2CU
##     0.6737644
```

Para poder calcular la matriz de confusión tendremos que determinar el óptimo cut-off

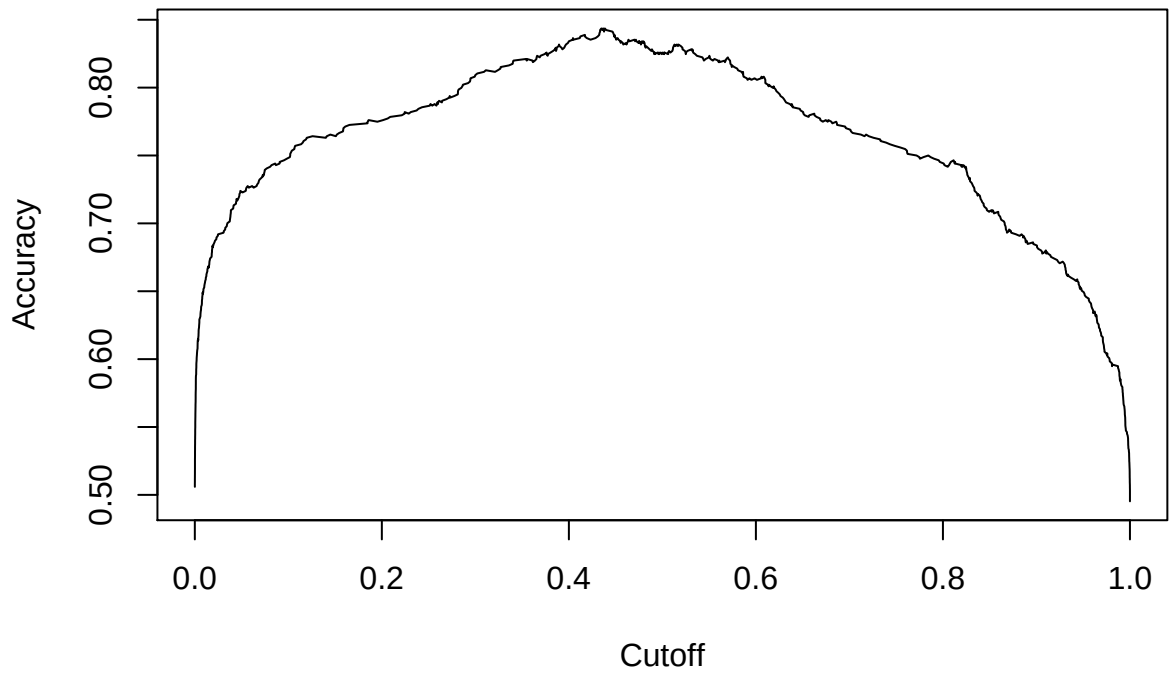
```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
prob <- predict(regresion_buena, datos_test_1, type = "response")
prediccion <- prediction(prob, datos_test_1$cat2)
eval <- performance(prediccion, "acc")
plot(eval)
```




```
max <- which.max(slot(eval, "y.values")[[1]])
acc <- slot(eval, "y.values")[[1]][max]
cutoff <- slot(eval, "x.values")[[1]][max]
print(c(Accuracy = acc, Cutoff = cutoff))
```

```
##    Accuracy Cutoff.2732
##   0.8436019   0.4384306
```

Con una precisión para nuestro modelo de un 84,36% obtenemos el mejor cut-off de un 0.4384306

```
logit_pred <- factor(prob > 0.4384306, levels = c(FALSE, TRUE), labels = c("Consumo de vacuno no bajo",
```

Creamos nuestra matriz de confusión, en donde relacionamos los valores actuales con las observaciones predichas

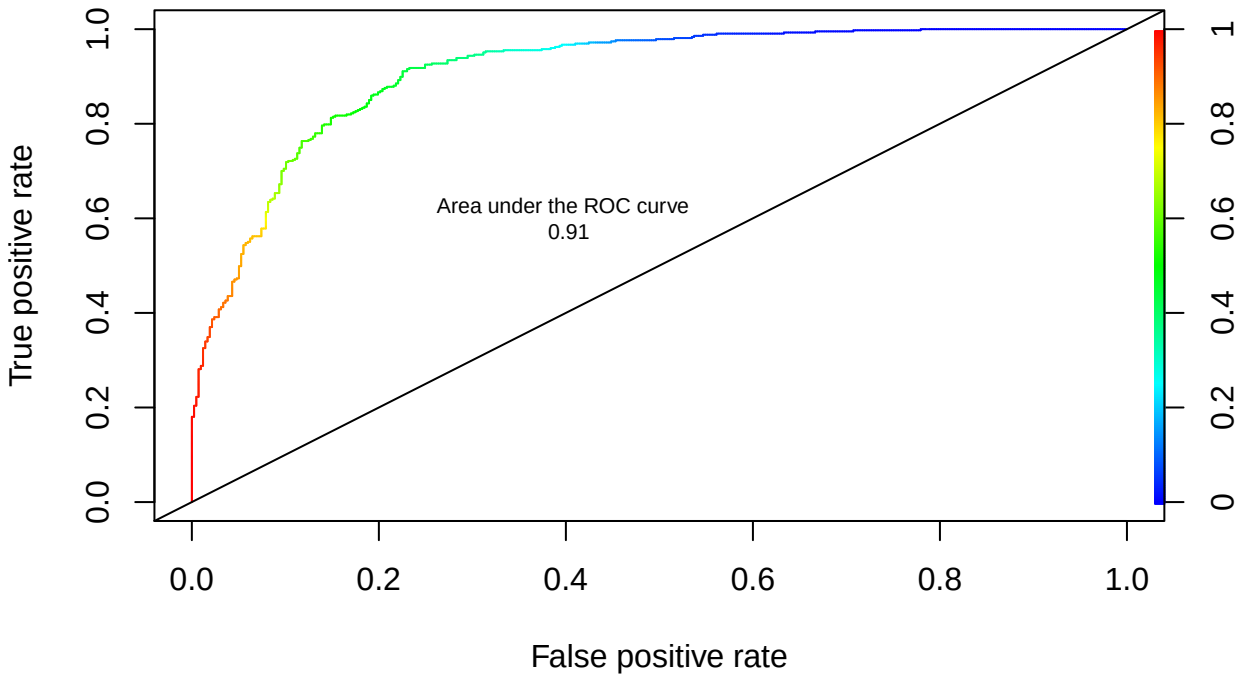
```
confution_table <- table(datos_test_1$cat2, logit_pred, dnn = c("Actual", "Predicted"))
confution_table
```

```
##       Predicted
## Actual Consumo de vacuno no bajo Consumo de vacuno bajo
##      0                      323                     94
##      1                       39                    388
```

Al analizar los resultados arrojados por la matriz de confusión se destaca que el modelo coloca como falsos positivos a 39 individuos, es decir, predice que estas personas estarán en el régimen consumo de carne de vacuno no bajo cuando en realidad están en régimen de consumir un porcentaje de carne vacuna anual bajo, y por otra parte el modelo coloca como falsos negativos (también conocido como error de tipo II) una cantidad de 94 individuos, con la misma interpretación que la anterior pero en sentido contrario.

Vamos a representar la curva ROC que relaciona la proporción de los falsos positivos con los verdaderos positivos

```
prediccion1 <- prediction(prob, datos_test_1$cat2)
AUC <- performance(prediccion1, "auc")
perf <- performance(prediccion1, "tpr", "fpr")
plot(perf, colorize = TRUE) # Establecemos el color.
abline(a = 0, b = 1)
text(0.4, 0.6, paste(AUC@y.name, "\n", round(unlist(AUC@y.values), 3)), cex = 0.7)
```



El área pode debajo de la curva ROC es de 0.91.Sabiendo que el AUC toma valores comprendidos entre 0 y 1, vemos que a través del modelo de regresión logística el rendimiento en cuanto a clasificación es bastante

notable.

ÁRBOLES DE CLASIFICACIÓN

Para su realización, se fundamenta en la elección de un modelo que busca el error de clasificación mínimo asociado a una determinada magnitud de la complejidad del árbol (parámetro de complejidad), una vez desarrollado el número máximo de nodos posibles.

Dichos indicadores que nos llevan a reflexionar que los determinantes en el consumo de carne vacuna estén asociadas a las variables independientes estadísticamente más significativas. Para emplear este ánalisis vamos a realizarlo con los árboles de clasificación tradicionales y con el árbol podado, para finalmente llegar a las conclusiones pertinentes.

ÁRBOL TRADICIONAL

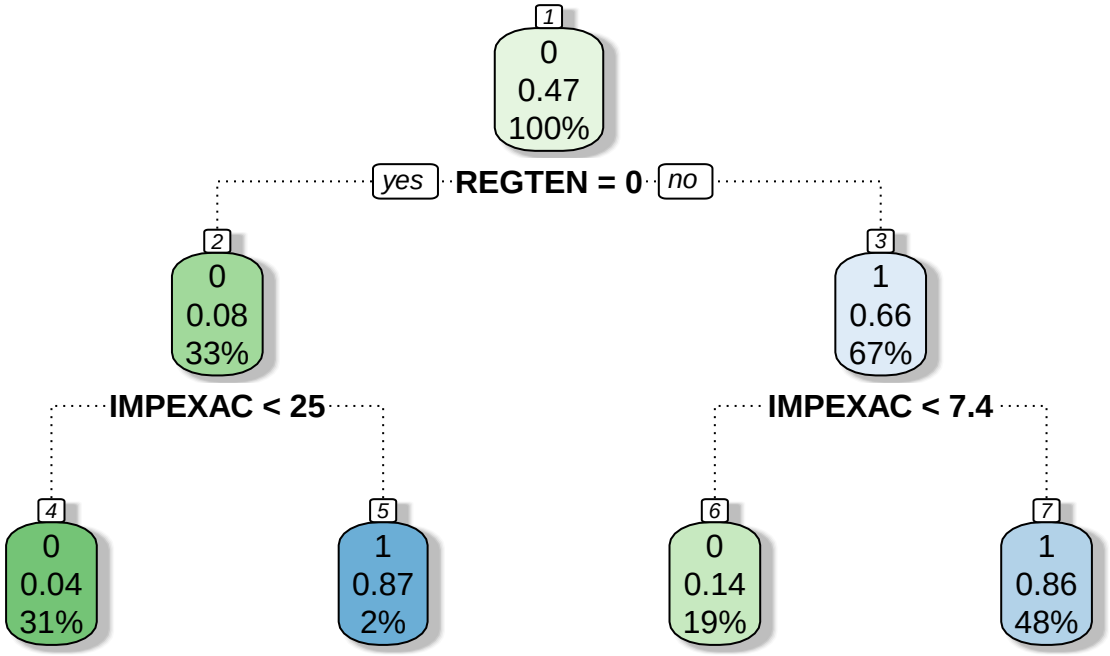
```
library(rpart)

set.seed(1234)
arbol <- rpart(cat2 ~ .,
               data=datos_train_1,
               method="class",
               parms=list(split="information"))
print(arbol)
```

```
## n= 3376
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 3376 1583 0 (0.53110190 0.46889810)
##   2) REGTEN=0 1102   92 0 (0.91651543 0.08348457)
##     4) IMPEXAC< 25.105 1040   38 0 (0.96346154 0.03653846) *
##     5) IMPEXAC>=25.105 62    8 1 (0.12903226 0.87096774) *
##   3) REGTEN=1 2274  783 1 (0.34432718 0.65567282)
##     6) IMPEXAC< 7.435 649   89 0 (0.86286595 0.13713405) *
##     7) IMPEXAC>=7.435 1625  223 1 (0.13723077 0.86276923) *
```

A continuación pasamos a representar el Árbol de clasificación

```
library(rpart.plot)
rpart.plot(arbol, box.palette = "GnBu", branch.lty = 3,
           shadow.col = "gray",
           nn = TRUE, main = "Árbol de clasificación por consumo de carne vacuna")
```

Árbol de clasificación por consumo de carne vacuna



A través del árbol de clasificación sin podar el primer criterio de clasificación es el régimen de tenencia, en donde podemos ver que aquellos que estén en el régimen de tenencia en propiedad un 67% de ellos consumen carne vacuna y sobre ellos un 48% con ingresos netos mensuales superiores a 7.4 (expresados como cientos de euros) son los que tienen un consumo cárnico. Tomado todo ello en su conjunto podemos decir que aquellos hogares que están en un régimen de tenencia en propiedad y con unos ingresos netos mensuales superiores a 7.4 son aquellos que tienen una mayor predilección de consumir producto vacuno.

La matriz de confusión del árbol tradicional sin podar arroja una precisión del 87.91469

```
arbol.pred1 <- predict(arbol, datos_test_1, type="class")

tabla.clasif.arbol1 <- table(datos_test_1$cat2, arbol.pred1,
                             dnn=c("Actual", "Predicted"))

tabla.clasif.arbol1
```

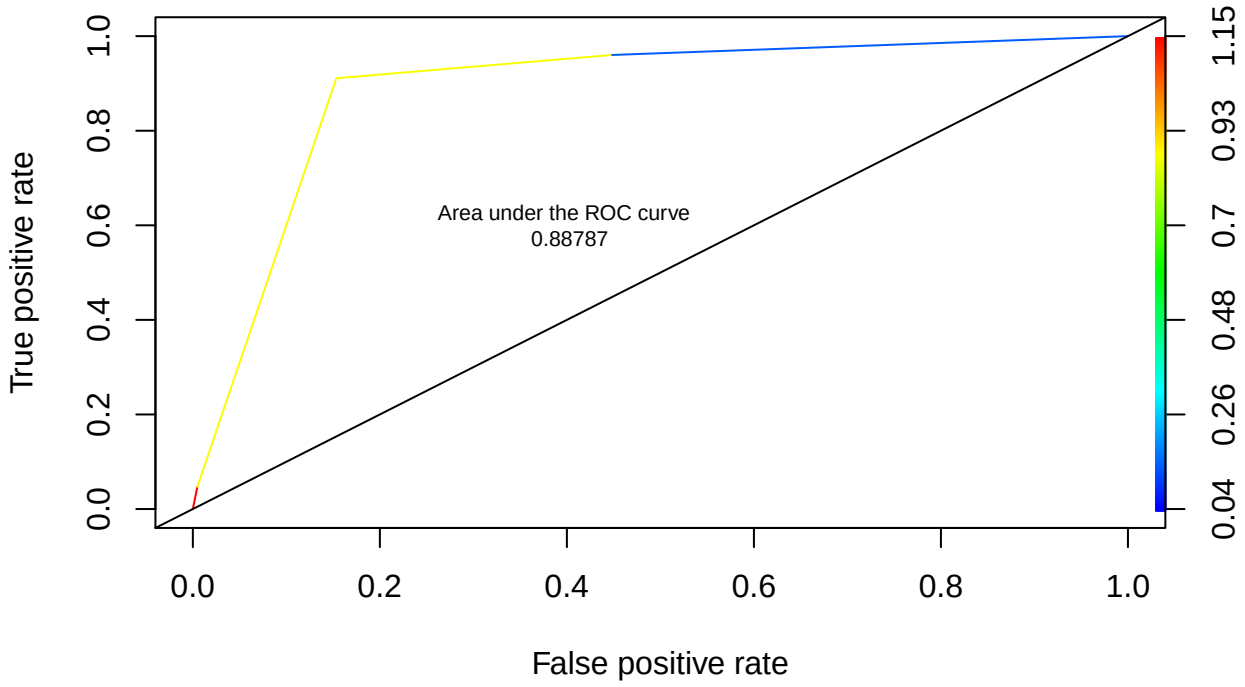
```
##        Predicted
## Actual   0   1
##      0 353  64
##      1  38 389
```

```
tcc2 <- 100 * sum(diag(tabla.clasif.arbol1))/sum(tabla.clasif.arbol1)
tcc2
```

```
## [1] 87.91469
```

Con la curva ROC el AUC es de 0.88787

```
prediccion_arbol <- predict(arbol, datos_test_1, type="prob")[,2]
pred_arbol = prediction(prediccion_arbol, datos_test_1$cat2)
AUC3 <- performance(pred_arbol, "auc")
perf3 <- performance(pred_arbol, "tpr", "fpr")
plot(perf3, colorize = TRUE)
abline(a = 0, b = 1)
text(0.4, 0.6, paste(AUC3@y.name, "\n", round(unlist(AUC3@y.values), 5)), cex = 0.7)
```



ÁRBOL PODADO

Determinamos el parámetro de complejidad relativo al error mínimo, y vemos que es de 0.01

```
arbol$cptable[which.min(arbol$cptable[,"xerror"]),"CP"]
```

```
## [1] 0.01
```

A través de este comando podemos determinar el xerror mínimo calculado, el cual es de 0.22868

```
printcp(arbol)
```

```
##
## Classification tree:
## rpart(formula = cat2 ~ ., data = datos_train_1, method = "class",
##     parms = list(split = "information"))
##
## Variables actually used in tree construction:
## [1] IMPEXAC REGTEN
##
```

```
## Root node error: 1583/3376 = 0.4689
##
## n= 3376
##
##         CP nsplit rel error  xerror     xstd
## 1 0.447252      0   1.00000 1.00000 0.018317
## 2 0.297536      1   0.55275 0.55275 0.016083
## 3 0.029059      2   0.25521 0.25711 0.011951
## 4 0.010000      3   0.22615 0.22868 0.011356
```

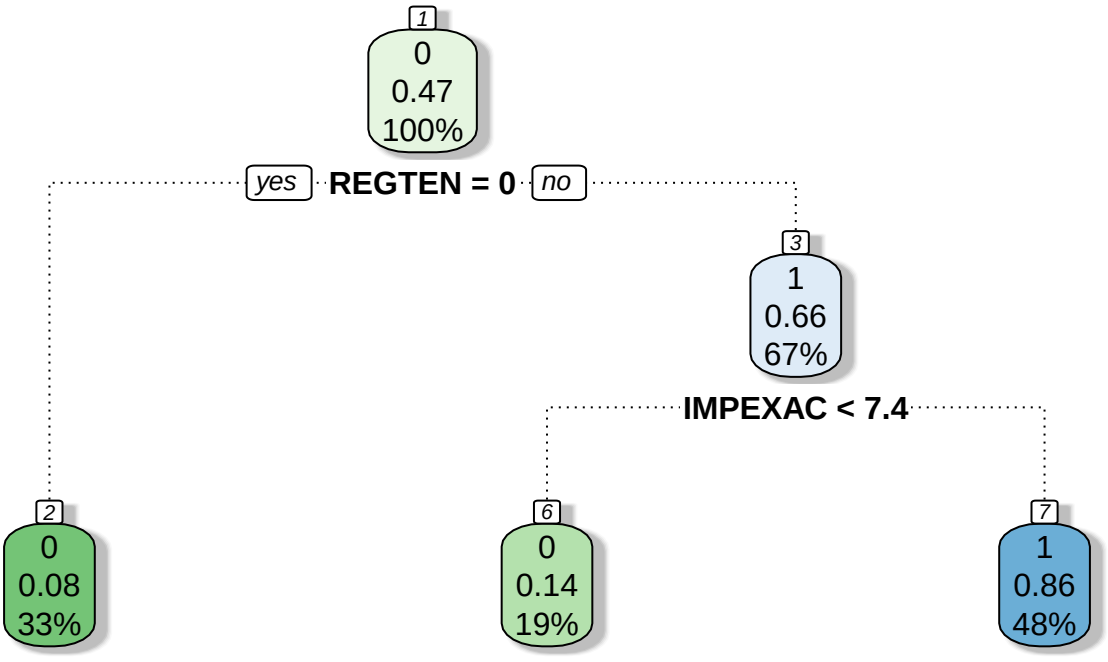
Con la finalidad de reducir el error de clasificación se realiza el podado del árbol de clasificación. Para este modelo el error relativo mínimo fue de 0.22868; sin embargo dado que la suma del error relativo y su desviación estándar ($0.22868 + 0.011356 = 0.240036$) fue mayor al error relativo que las antecede, se opta por emplear dicho error asociado a una criterio de complejidad 0.029059; lo cual conlleva que el núméto total de divisiones sea de una división.

Realizamos la poda entonces con el cp= 0.029059

```
arbol_podado = prune(arbol, cp = 0.029059)
```

Representamos el árbol de clasificación podado

```
rpart.plot(arbol_podado, box.palette = "GnBu", branch.lty = 3,
           shadow.col = "gray",
           nn = TRUE, main = "Árbol de clasificación para el consumo de carne vacuna")
```



Con el árbol podado aquellos hogarres que estén en régimen de tenencia en alquiler consumen un 33% carne vacuna. Sin embargo, para aquellos hogares que están en régimen de tenencia en propiedad un 67% de estos con ingresos netos superiores a 7.4 (en cientos de euros) consumen un 48% de carne vacuna.

Realizamos a continuación nuestra matriz de confusión para comparar los resultados obtenidos, el cual para el árbol podado obtenemos un accuracy de 85.78199

```
arbol_prediccion <- predict(arbol_podado, datos_test_1, type = "class")

#  Se trabaja sobre el arbol podado
arbol_resultado_total <- table(datos_test_1$cat2, arbol_prediccion,
                               dnn = c("Actual", "Predicted"))

# Tabla de doble entrada
arbol_resultado_total
```

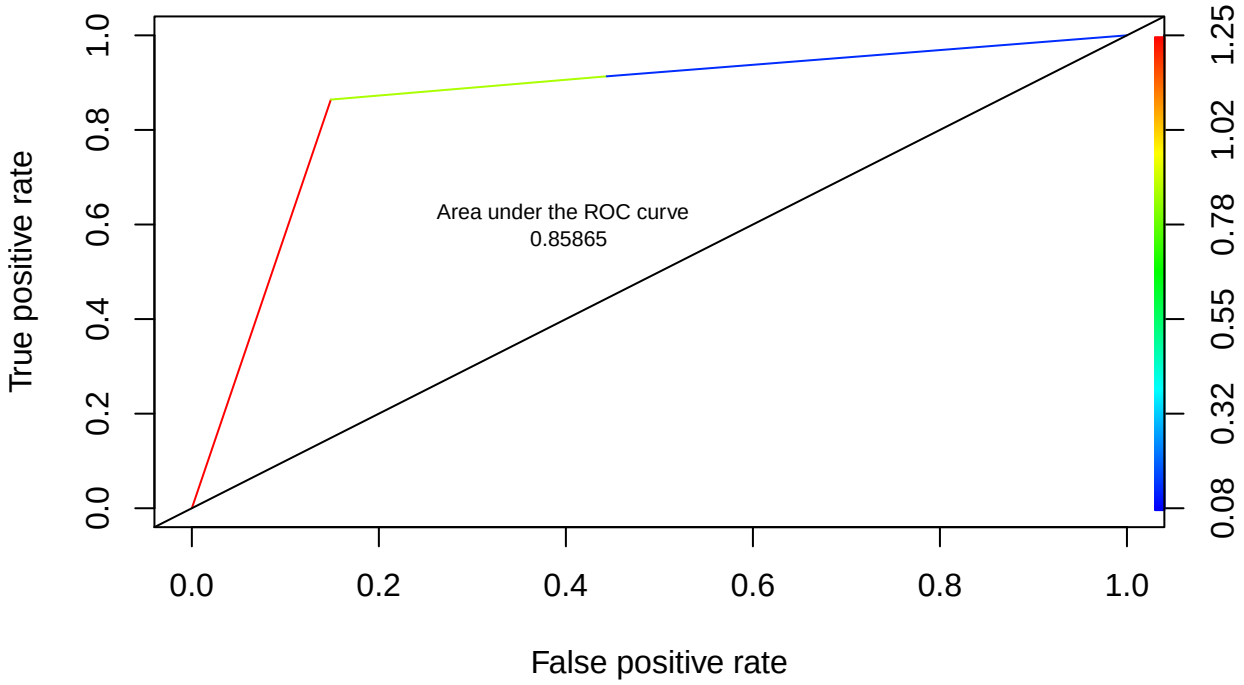
```
##       Predicted
## Actual   0   1
##      0 355  62
##      1  58 369
```

```
tcc1 <- 100 * sum(diag(arbol_resultado_total))/sum(arbol_resultado_total)
tcc1
```

```
## [1] 85.78199
```

Realizando la representación de la curva ROC tenemos un área po debajo de la curva de 0.887

```
prediccion3 <- predict(arbol_podado, datos_test_1, type="prob")[,2]
pred3 = prediction(prediccion3, datos_test_1$cat2)

AUC5 <- performance(pred3, "auc")

perf5 <- performance(pred3, "tpr", "fpr")

plot(perf5, colorize = TRUE)

abline(a = 0, b = 1)

text(0.4, 0.6, paste(AUC5@y.name, "\n", round(unlist(AUC5@y.values), 5)), cex = 0.7)
```



CONCLUSIONES EXTRAÍDAS SEGÚN LA REGRESIÓN LOGÍSTICA Y LOS ARBOLES DE CLASFICACIÓN

Table 2: Tabla de que muestra la precisión y la curva ROC para la regresión logística y árbol de clasificación

	Regresión Logística	Árbol_tradicional	Árbol_podado
Curva ROC	0.91	0.88	0.85
Accuracy	84.3%	87.91%	85.78%

La mejor precisión para el consumo de CAT2 la otorga el árbol de clasificación tradicional, con una accuracy del 87.91%, en segundo lugar sería el árbol podado y por último la regresión logística. Sin embargo, según la curva ROC, la mejor área por debajo de la curva es la regresión logística, en segundo lugar el árbol tradicional y por último, el árbol podado.

Tomando todo ello en su conjunto considero que el mejor modelo para predecir la variable dependiente CAT2 es a través del árbol tradicional, en segundo lugar, a través de la regresión logística y por último a través del árbol podado.

PREGUNTA 3: Aplicar un árbol de clasificación para la variable CAT3

Emplearemos la variable CAT3 como variable predictora/variable dependiente para el análisis de los árboles de clasificación y eliminaremos de nuestro modelo a la variable CAT2 para evitar incurrir en un problema de

multicolinealidad y prevenir posibles problemas de sobreajuste, puesto que las dos explicar el consumo de carne vacuno anual, aunque en diferentes niveles. CAT2 está factorizada en dos niveles (binomial) y CAT3 está factorizada en 3 niveles (multivariada).

Emplearemos la misma secuencia que en la pregunta 2, es decir, en primer lugar realizaremos el análisis con el árbol tradicional y en segundo lugar emplearemos el árbol podado

ÁRBOL TRADICIONAL

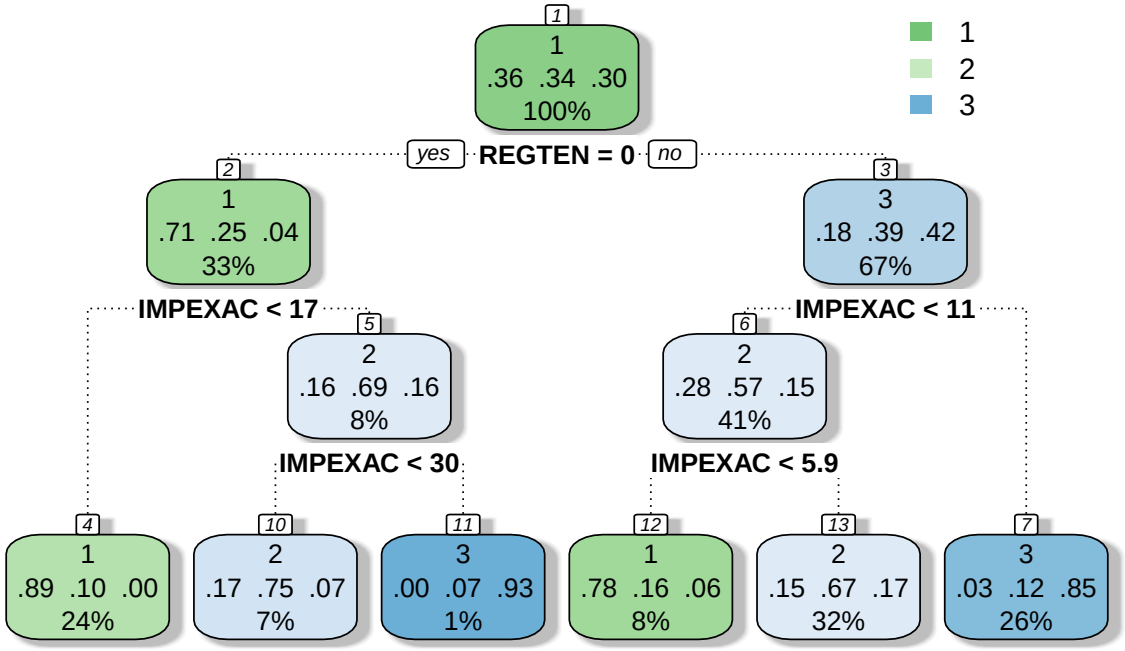
```
datos_train_2 <- datos_train[,-10]
datos_test_2 <- datos_test[,-10]
```

```
set.seed(1234)
arbol1 <- rpart(cat3 ~ .,
                data=datos_train_2,
                method="class",
                parms=list(split="information"))
print(arbol1)
```

```
## n= 3376
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 3376 2174 1 (0.35604265 0.34478673 0.29917062)
##    2) REGTEN=0 1102  320 1 (0.70961887 0.24863884 0.04174229)
##      4) IMPEXAC< 16.97 827   88 1 (0.89359129 0.10278114 0.00362757) *
##      5) IMPEXAC>=16.97 275   86 2 (0.15636364 0.68727273 0.15636364)
##       10) IMPEXAC< 30.435 248   61 2 (0.17338710 0.75403226 0.07258065) *
##       11) IMPEXAC>=30.435 27    2 3 (0.00000000 0.07407407 0.92592593) *
##    3) REGTEN=1 2274 1310 3 (0.18469657 0.39138083 0.42392260)
##      6) IMPEXAC< 11.445 1383  597 2 (0.28271873 0.56832972 0.14895155)
##       12) IMPEXAC< 5.94 286   64 1 (0.77622378 0.16083916 0.06293706) *
##       13) IMPEXAC>=5.94 1097  357 2 (0.15405652 0.67456700 0.17137648) *
##      7) IMPEXAC>=11.445 891  133 3 (0.03254770 0.11672278 0.85072952) *
```

```
library(rpart.plot)
rpart.plot(arbol1, box.palette = "GnBu", branch.lty = 3,
           shadow.col = "gray",
           nn = TRUE, main = "Árbol de clasificación por consumo de carne vacuna")
```

Árbol de clasificación por consumo de carne vacuna



A través del árbol de clasificación sin podar el primer criterio de clasificación es el régimen de tenencia, en donde un 67% de los hogares que están en régimen de tenencia en propiedad tienen un consumo medio-alto de carne vacuna (categoría_3) y sobre ellos aquellos que tienen unos ingresos netos mensuales superiores a 11 (expresado como cientos de euros) un 41% tienen un mayor consumo de producto vacuno. El último criterio de clasificación lo establece también los ingresos netos mensuales pero con el umbral del 5.9, en donde de ese 41%, el 8% de los hogares que tienen una renta mensual superior al umbral citado tendrán mayor predilección de consumir este tipo de producto.

La matriz de confusión del árbol tradicional sin podar arroja una precisión del 76.65877

```
arbol.pred2 <- predict(arbol1, datos_test_2, type="class")

tabla.clasif.arbol2 <- table(datos_test_2$cat3, arbol.pred2,
                             dnn=c("Actual", "Predicted"))

tabla.clasif.arbol2
```

```
##        Predicted
## Actual   1   2   3
##      1 200  59  11
##      2  40 239  30
##      3   7  50 208
```

```
tcc3 <- 100 * sum(diag(tabla.clasif.arbol2))/sum(tabla.clasif.arbol2)
tcc3
```

```
## [1] 76.65877
```

No podemos representar la curva ROC para la variable dependiente CAT3 ya que solo permite la representación

gráfica para variables binarias

ÁRBOL PODADO

Determinamos el parámetro de complejidad relativo al error mínimo, y vemos que es de 0.01

```
arbol1$cptable[which.min(arbol1$cptable[,"xerror"]),"CP"]
```

```
## [1] 0.01
```

A través de este comando podemos determinar el xerror mínimo calculado, el cual es 0.34499

```
printcp(arbol1)
```

```
##
## Classification tree:
## rpart(formula = cat3 ~ ., data = datos_train_2, method = "class",
##     parms = list(split = "information"))
##
## Variables actually used in tree construction:
## [1] IMPEXAC REGTEN
##
## Root node error: 2174/3376 = 0.64396
##
## n= 3376
##
##         CP nsplit rel error  xerror     xstd
## 1 0.258510      0   1.00000 1.00000 0.012797
## 2 0.080957      2   0.48298 0.48436 0.012382
## 3 0.067157      3   0.40202 0.42778 0.011940
## 4 0.010580      4   0.33487 0.36983 0.011384
## 5 0.010000      5   0.32429 0.34499 0.011110
```

Con la finalidad de reducir el error de clasificación se realiza el podado del árbol de clasificación. Para este modelo el error relativo mínimo fue de 0.34499; sin embargo dado que la suma del error relativo y su desviación estándar (0.34499 + 0.011110 = 0.3561) fue mayor al error relativo que las antecede, se opta por emplear dicho error asociado a una criterio de complejidad 0.010580; lo cual conlleva que el núméto total de divisiones sea de una división.

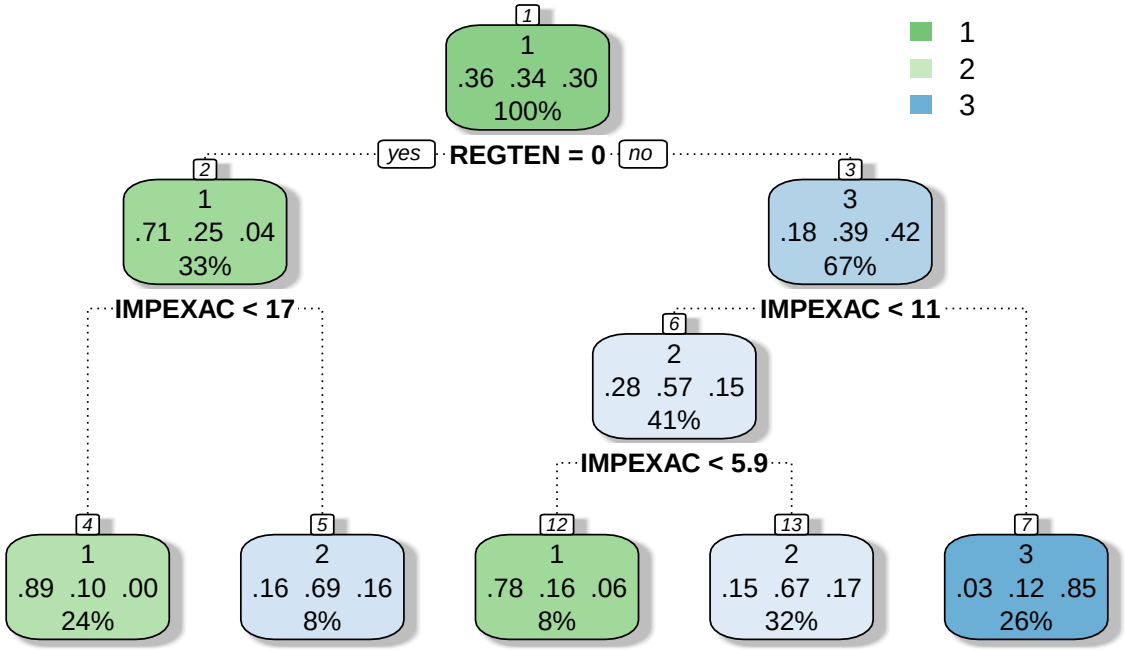
Realizamos la poda entonces con el cp= 0.010580

```
arbol_podado1 = prune(arbol1, cp = 0.010580)
```

Representamos gráficamente nuestro árbol podado

```
rpart.plot(arbol_podado1, box.palette = "GnBu", branch.lty = 3,
           shadow.col = "gray",
           nn = TRUE, main = "Árbol de clasificación para el consumo de carne vacuna")
```

Árbol de clasificación para el consumo de carne vacuna



A través del árbol de clasificación podado el primer criterio de clasificación es el régimen de tenencia, en donde un 67% de los hogares que están en régimen de tenencia en propiedad tienen un consumo medio-alto de carne vacuna (categoría_3) y sobre ellos aquellos que tienen unos ingresos netos mensuales superiores a 11 (expresado como cientos de euros) un 41% tienen un mayor consumo de producto vacuno. El último criterio de clasificación lo establece también los ingresos netos mensuales pero con el umbral del 5.9, en donde de ese 41%, el 8% de los hogares que tienen una renta mensual superior al umbral citado tendrán mayor predilección de consumir este tipo de producto.

Realizamos a continuación nuestra matriz de confusión para comparar los resultados obtenidos, el cual para el árbol podado obtenemos un accuracy de 75.7109

```
arbol_prediccion1 <- predict(arbol_podado1, datos_test_2, type = "class")

#  Se trabaja sobre el arbol podado
arbol_resultado_total1 <- table(datos_test_2$cat3, arbol_prediccion1,
                                dnn = c("Actual", "Predicted"))

# Tabla de doble entrada
arbol_resultado_total1
```

```
##        Predicted
## Actual   1   2   3
##      1 200  59  11
##      2  40 240  29
##      3   7  59 199
```

```
tcc2 <- 100 * sum(diag(arbol_resultado_total1))/sum(arbol_resultado_total1)
tcc2
```

```
## [1] 75.7109
```

No podemos representar la curva ROC para la variable dependiente CAT3 ya que solo permite la representación gráfica para variables binarias

CONCLUSIONES EXTRAÍDAS SEGÚN LOS ARBOLES DE CLASFICACÍÓN PARA CAT3

Table 3: Conclusiones para los árboles de clasificación empleando CAT3

	Árbol_tradicional	Árbol_podado
Curva ROC	No tiene al no ser binaria	No tiene al no ser binaria
Accuracy	76.65877	75.7109

Observamos que los resultados arrojados a la hora de realizar la predicción clasificadora para la variable CAT3 son mejores sin realizar lapoda, por lo que el árbol tradicional sin aplicar la poda resulta mucho mejor que el árbol podado para realizar la predicción de clasificación para la variable CAT3.

PREGUNTA4. COMPARACIÓN DE RESULTADOS Y CONCLUSIONES FINALES

Table 4: Conclusiones finales según los modelos

	Regres. Logística_CAT2	Árbol_sinpoda_CAT2	Árbol_podado_CAT2	Árbol_sinpoda_CAT3	Árbol_podado_CAT3
Curva ROC	0.91	0.88	0.85	No tiene	No tiene
Accuracy	84,3%	87.91%	85.78%	76.65%	75.71%

Tomado todo ello en su conjunto podemos ver que tanto a través del método de clasificación de regresión logística como para los árboles de clasificación existe una mejor precisión en cuanto a la predicción de clasificación de la variable CAT2 frente a CAT3.

Nota*: Tener en cuenta que variables no binomiales como CAT3 no pueden ser representadas a través de la curva ROC.

ANEXO

ANEXOS en donde evaluamos la importancia que tienen cada una de las variables predictoras

- Por una parte, con la variable depediente CAT2 frente al resto de variables

```
library(tidyverse)
```

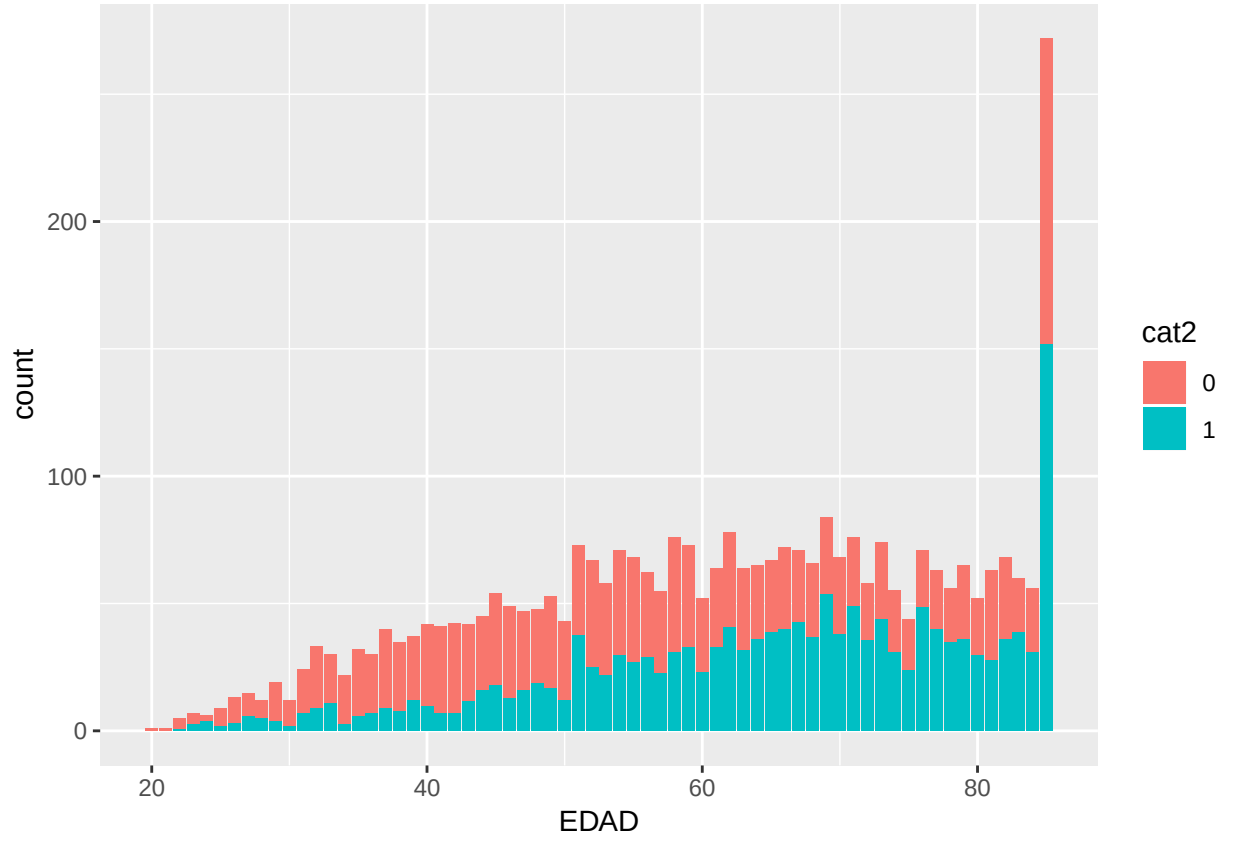
```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v stringr 1.3.1
## v readr   1.3.0     v forcats 0.3.0
```

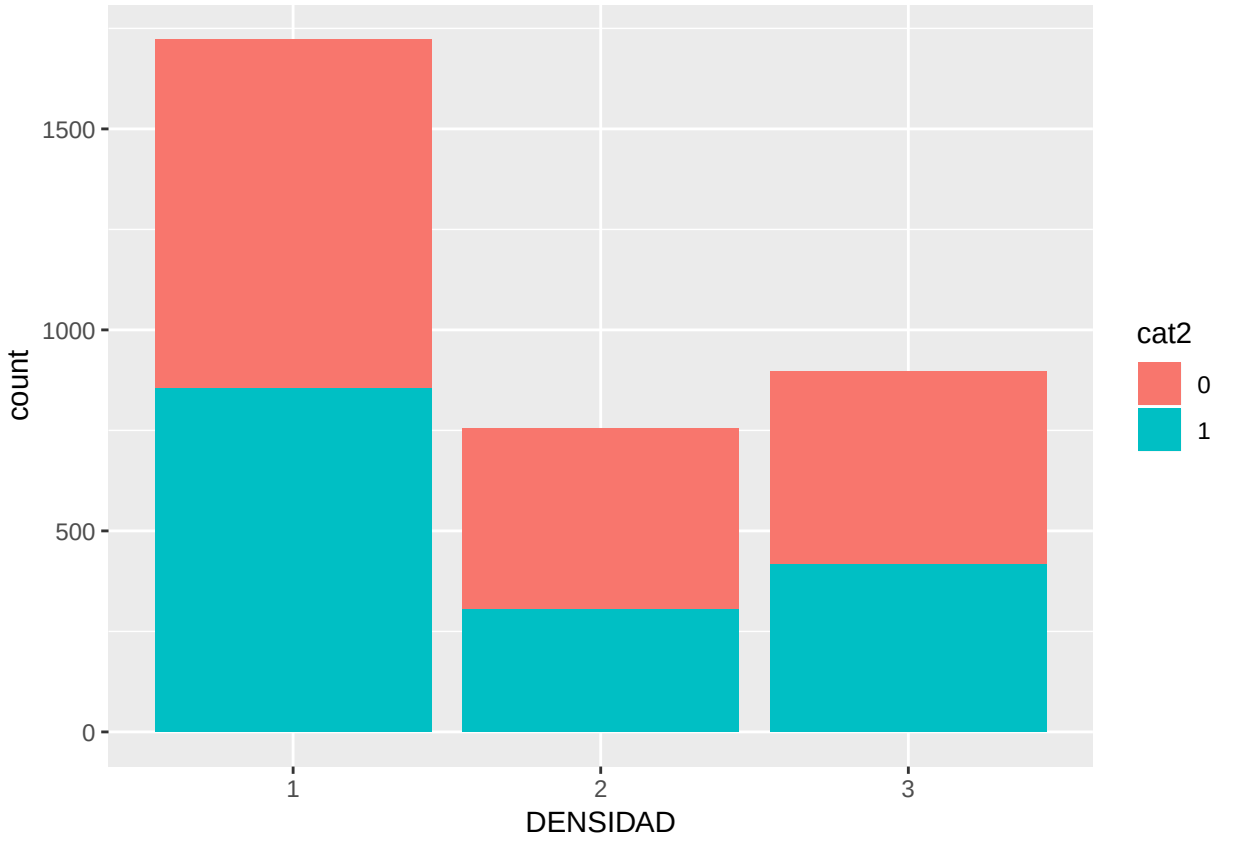
```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x MASS::select()  masks dplyr::select()
```

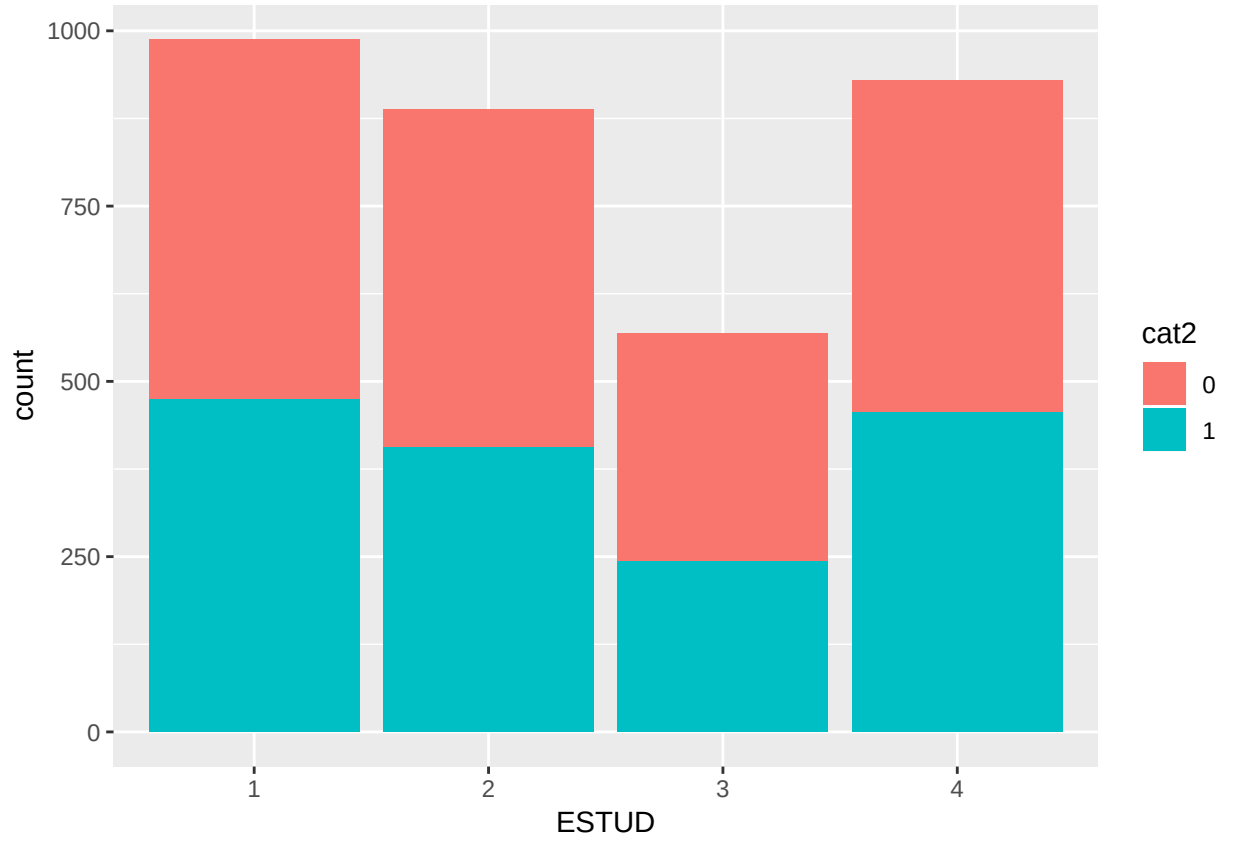
```
ggplot(data = datos_train_1, mapping = aes(x =EDAD )) +
  geom_bar(mapping = aes(fill = cat2))
```



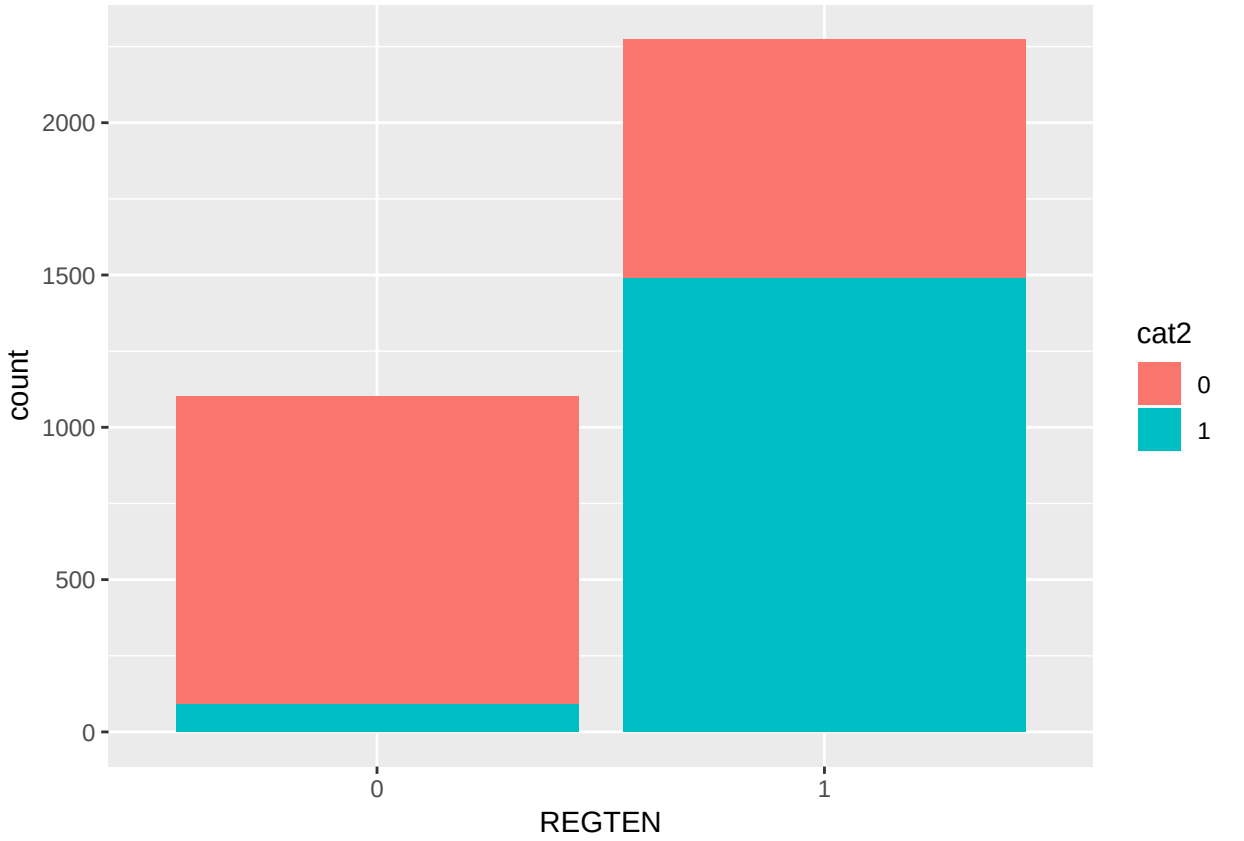

```
ggplot(data = datos_train_1, mapping = aes(x =DENSIDAD )) +
  geom_bar(mapping = aes(fill = cat2))
```



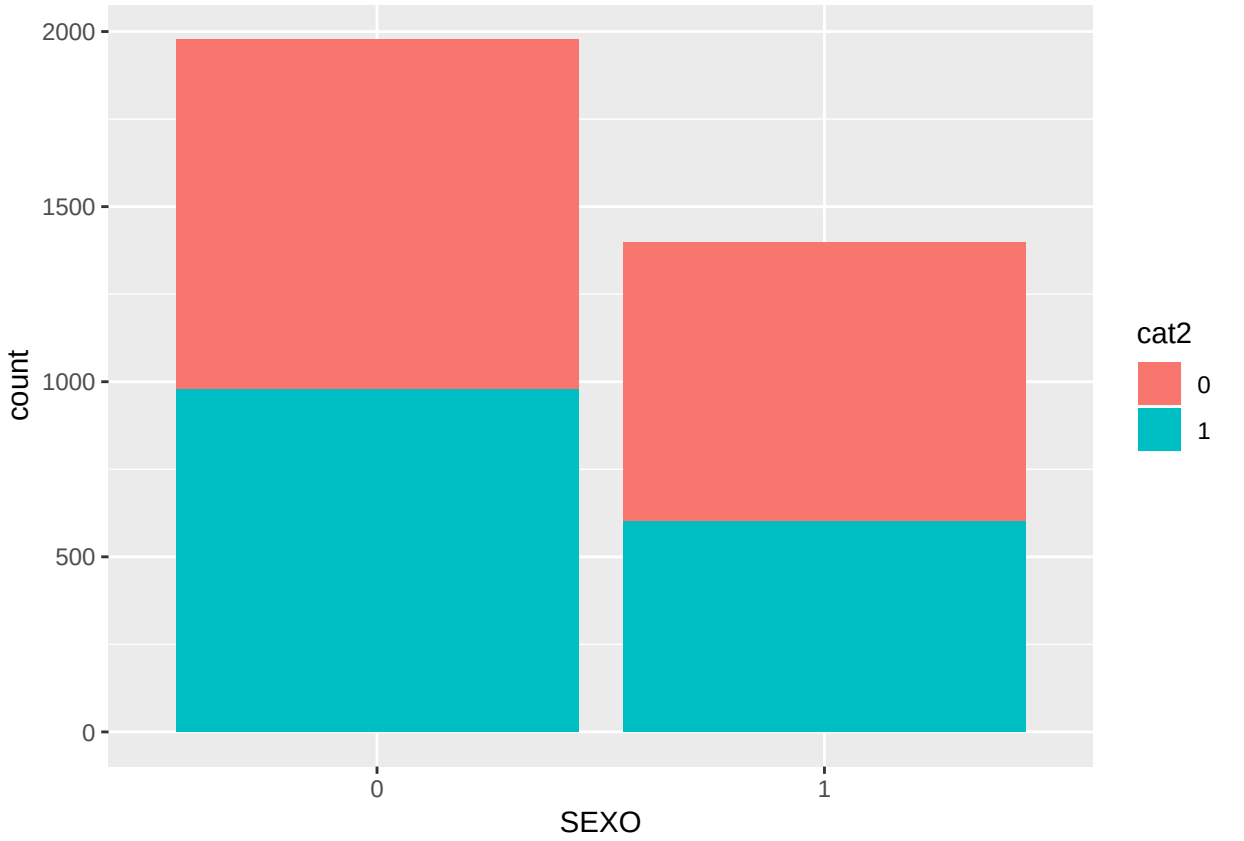

```
ggplot(data = datos_train_1, mapping = aes(x =ESTUD )) +
  geom_bar(mapping = aes(fill = cat2))
```




```
ggplot(data = datos_train_1, mapping = aes(x = REGTEN)) +
  geom_bar(mapping = aes(fill = cat2))
```




```
ggplot(data = datos_train_1, mapping = aes(x =SEXO )) +
  geom_bar(mapping = aes(fill = cat2))
```



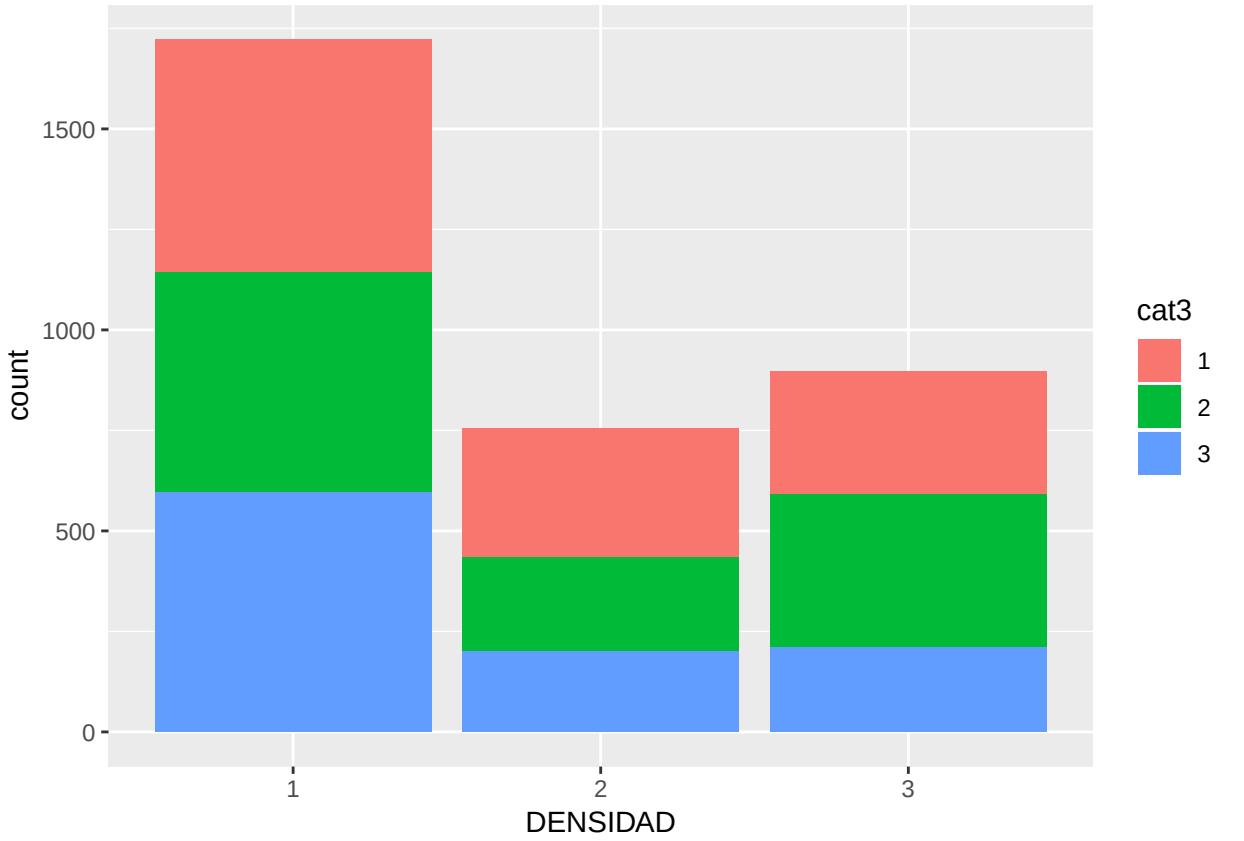
Conclusiones extraídas en base a los gráficos:

- Conforme pasan los años más hogares son los que tienen una predilección mayor por consumir carne vacuna.
- En las zonas densamente pobladas existe casi la misma proporción de hogares que tienen y no tienen un consumo bajo de carne vacuna anual. Sim embargo, para las zonas donde no existe tanta densidad de población (categoría2) los hogares prefieren consumir menos producto vacuno.
- Según el nivel de estudios completados, las diferencias más significativas están en la catgoría3, es decir, para aquellos individuos que han completado la segunda etapa de educación secundaria, en donde dichas personas prefieren no consumir tanta carne vacuna.
- Aquellas personas que están en régimen de alquiler consumen menos carne. En cambio, aquellas que están en régimen de tenencia en propiedad prefieren tener un porcentaje de consumo de carne vacuna mucho más acentuado.
- Los hombres tienen un nivel de consumo de carne vacuna inferior, comparándola con las mujeres.

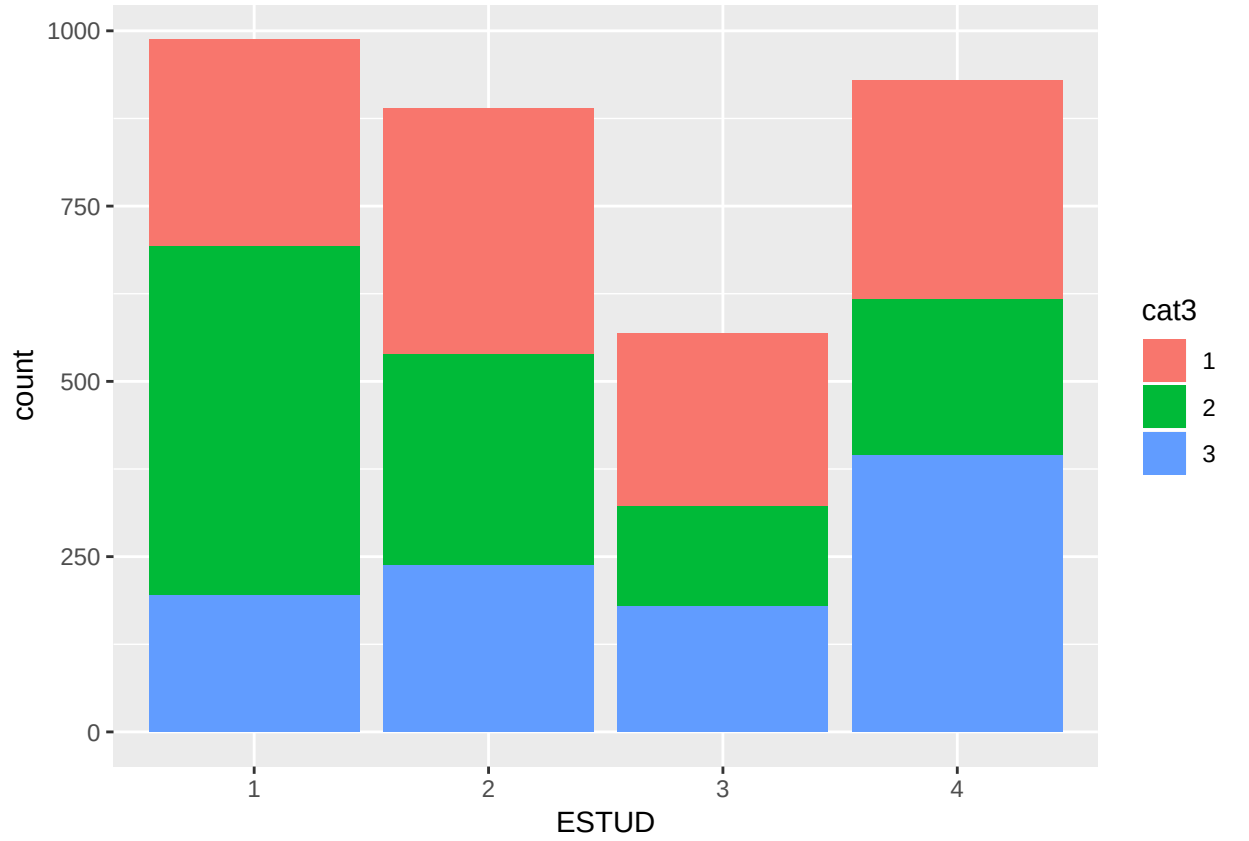
Por una parte, con la variable depediente CAT3 frente al resto de variables

```
ggplot(data = datos_train_2, mapping = aes(x =EDAD )) +
  geom_bar(mapping = aes(fill = cat3))
```

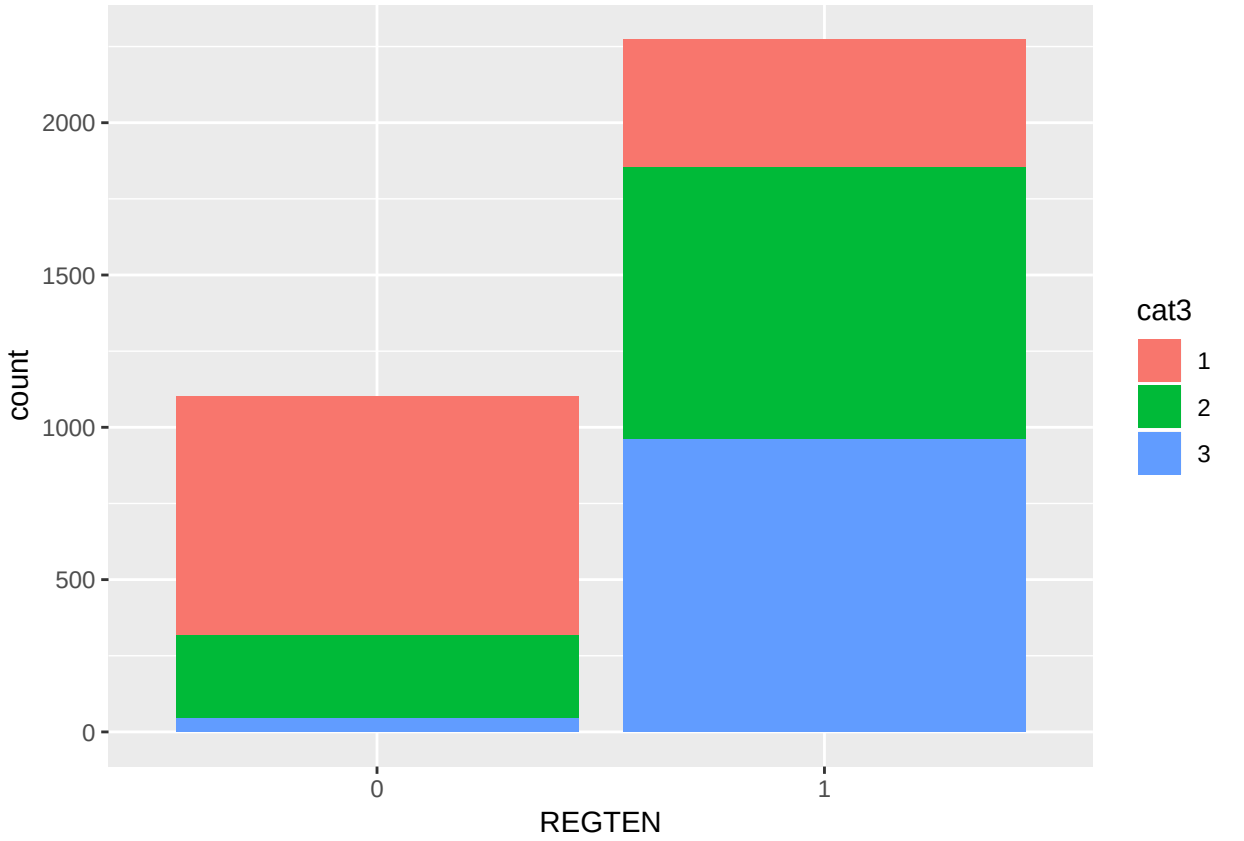

```
ggplot(data = datos_train_2, mapping = aes(x =DENSIDAD )) +
  geom_bar(mapping = aes(fill = cat3))
```



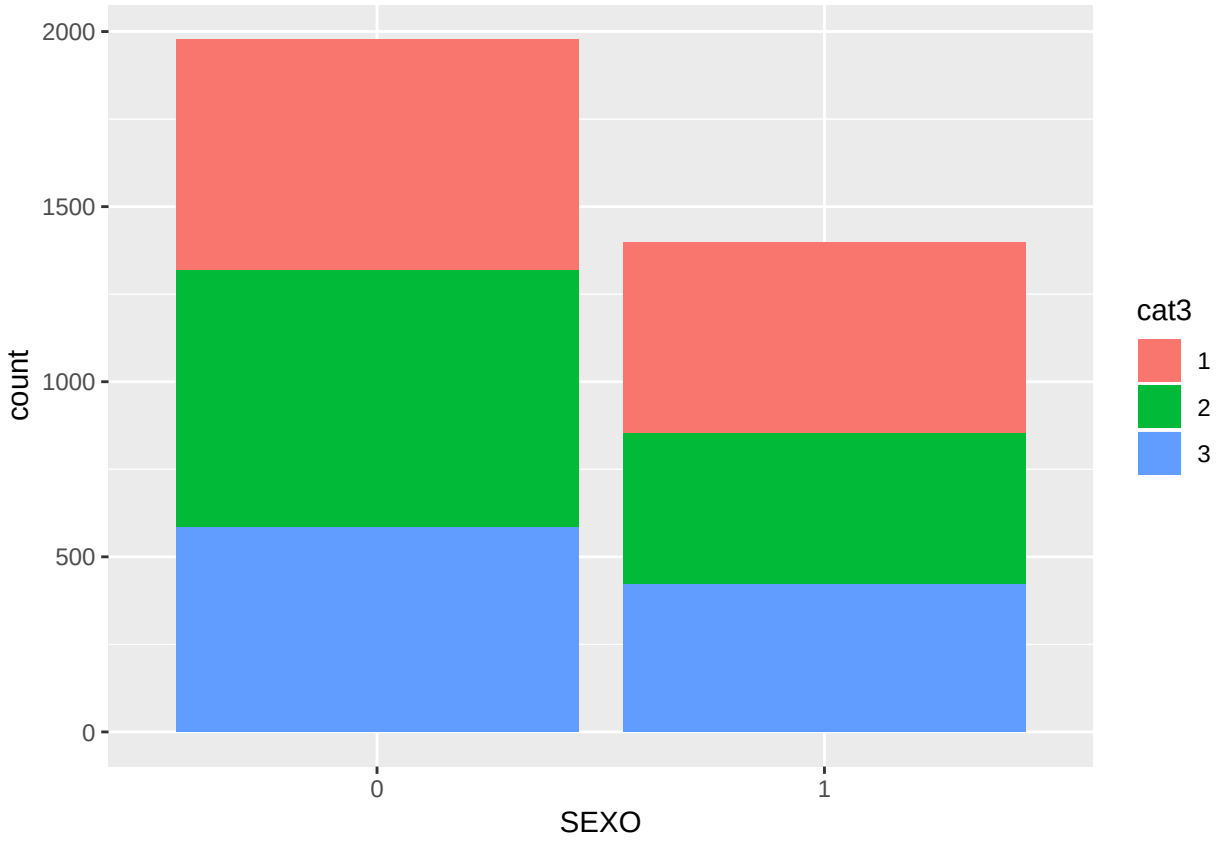

```
ggplot(data = datos_train_2, mapping = aes(x =ESTUD )) +
  geom_bar(mapping = aes(fill = cat3))
```




```
ggplot(data = datos_train_2, mapping = aes(x = REGTEN)) +
  geom_bar(mapping = aes(fill = cat3))
```




```
ggplot(data = datos_train_2, mapping = aes(x =SEXO )) +
  geom_bar(mapping = aes(fill = cat3))
```



Conclusiones extraídas en base a los gráficos:

- A medida que aumentan los años el consumo de carne vacuna va siendo más moderado.
- Tanto en poblaciones con alta densidad de población como aquellas intermedias existe una proporción entre hogares con consumo bajo, intermedio y alto de producto vacuno. Las diferencias radican en las zonas diseminadas en donde el consumo intermedio de carne es más acentuado.
- Aquellas personas que tienen un nivel de educación superior son aquellas que consumen más carne vacuna frente a aquellas que tienen un nivel bajo de estudios.
- Aquellas personas que estén en régimen de tenecia en alquiler son aquellas que tienen un consumo de carne muy bajo. Sim embargo, aquellas que tengan una propiedad tienen mayor inclinación a consumir este tipo de producto.
- Las mujeres tienen una inclinación a consumir moderamente la carne vacuna. En cambio, los hombres tienen una proporción mucho más marcada entre aquellos que consumen mucha carne vacuna frente a los que no tanta.