ProyectoFinal_MDD

Susan Valda - Andrew Serrano

2022-11-18

Contents

1 Introducción

2 Motivación

3 Marco Teórico

4 Metodología

Para investigar las características asociadas a la práctica del trabajo infantil se tomó como referencia la metodología KDD, por lo cual se plantean los siguientes pasos: 1. Recolección de la información referente

a encuestas que contemplen la situación de niños/niñas/adolescentes que realiacem algún tipo de trabajo infantil en Bolivia. 2. Selección y tratamiento de los datos. 3. Utilizar la librería RPART y CART de Rstudio para generar un modelo y árbol de decisión en base a los datos seleccionados. 5. Finalmenta se efectúa el análisis e interpretación del árbol de desición obtenido.

5 Datos

Se emplearon los datos de la Encuesta de niñas, niños y adolescentes (ENNA) que realizan una actividad laboral o trabajan 2016 obtenida del Instituto Nacional de Estadística (INE).

La base de datos ENNA cuenta con 10488 observaciones y 212 Variables, de las cuales se utilizarán las siguientes variables para la elaboración del árbol de decisión.

Código	Variable
depto	Departamento
area	Urbana Rural
ns001a_02	Sexo
ns001a_03	¿Cuántos años cumplidos tienes?
ns01a_01	¿Sabes leer y escribir?
ns01a_02a	¿Cuál fue el último NIVEL Y CURSO más alto que aprobaste? NIVEL O CICLO
ns01a_02b	¿Cuál fue el último NIVEL Y CURSO más alto que aprobaste?CURSO O GRADO
ns01a_03	Durante este año ¿Estás o estuviste inscrito en algún curso o grado de educación escolar, alternativa o superior?
ns01a_05c	¿En qué turno te has inscrito este año (2016)?
ns01a_06	¿Asistes/Asististe regularmente al curso al que te has inscrito este año (2016)?
ns03a_04	¿Estás de acuerdo con realizar estas tareas domésticas?
ns04a_02	Durante la semana pasada, ¿consideras que tuviste un tiempo adecuado para descanso(relajación, ocio sano) o recreación...?
ns03a_01a	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°1)
ns03a_01b	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°2)
ns03a_01c	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°3)
ns03a_01d	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°4)
ns03a_01e	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°5)
ns03a_01f	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°6)
ns03a_01g	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°7)
ns03a_01h	Durante la semana pasada, ¿realizaste para este hogar, alguna de las tareas domésticas indicadas a continuación? (Tarea N°8)
ns03a_03a	Durante la semana pasada, ¿realizaste estas tareas usualmente...?

La variable que ayudará a entrenar el modelo es:

Código	Variable
condac	Condición de actividad

5.1 Recopilación de la fuente de datos

5.1.1 Base de datos

```
library(haven)#para importar bases de datos

#Carga de datos
load(url("https://github.com/AlvaroLimber/EST-384/blob/master/data/nna16.RData?raw=true"))
```

5.2 Selección

5.3 Pre-procesado

```
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)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
bd <- nna %>% select(condac,depto,area,sexo=ns001a_02,edad=ns001a_03,sabeLeer=ns01a_01,
maxNivelAprobado=ns01a_02a,maxCursoAprobado=ns01a_02b,inscAlternativa=ns01a_03,inscrito=ns01a_05c,
asistencia=ns01a_06,tareasDomesticas=ns03a_04,tiempoDescanso=ns04a_02,Tarea1=ns03a_01a,Tarea2=ns03a_01b
Tarea3=ns03a_01c,Tarea4=ns03a_01d,Tarea5=ns03a_01e,Tarea6=ns03a_01f,Tarea7=ns03a_01g,
Tarea8=ns03a_01h,usual=ns03a_03a)
```

5.3.1 Datos pre-procesados

5.4 Transformación

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x Hmisc::src()       masks dplyr::src()
## x Hmisc::summarize() masks dplyr::summarize()
```

```
#Omitir los datos vacíos o nulos.
bd <- bd %>% na.omit(bd)
```

5.5 Minería de datos

```
## Elaboración del modelo
library(rpart)
set.seed(123)
index = sample(1:2, nrow(bd), replace = TRUE, prob=c(0.7, 0.3))
prop.table(table(index))
```

```
## index
##         1         2
## 0.7032219 0.2967781
```

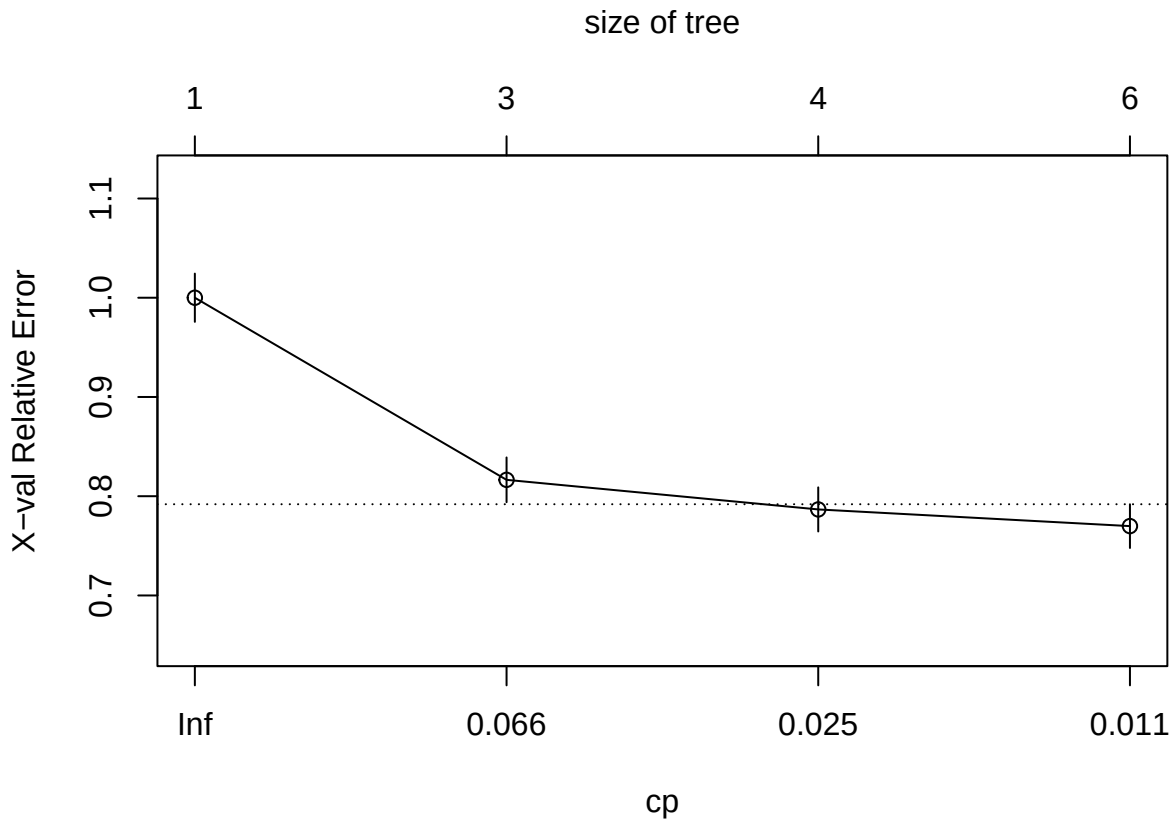
```
trainbd<-bd[index==1,]
testbd<-bd[index==2,]
mod1<-rpart(condac~.,data=trainbd)

#Explorar los nodos creados por rpart
mod1
```

```
## n= 5784
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 5784 1308 Sin actividad laboral o trabajo (0.7738589 0.2261411)
##    2) area=Urbana 4219  560 Sin actividad laboral o trabajo (0.8672671 0.1327329) *
##    3) area=Rural 1565  748 Sin actividad laboral o trabajo (0.5220447 0.4779553)
##      6) depto=Cochabamba,Oruro,Santa Cruz,Beni,Pando 713  200 Sin actividad laboral o trabajo (0.719
##      7) depto=Chuquisaca,La Paz,Potosí,Tarija 852  304 Con actividad laboral o trabajo (0.3568075 0.
##       14) Tarea7=2.No 331  135 Sin actividad laboral o trabajo (0.5921450 0.4078550)
##         28) edad< 8.5 113   19 Sin actividad laboral o trabajo (0.8318584 0.1681416) *
##         29) edad>=8.5 218  102 Con actividad laboral o trabajo (0.4678899 0.5321101)
##           58) depto=Chuquisaca,Potosí,Tarija 154   67 Sin actividad laboral o trabajo (0.5649351 0.43
##           59) depto=La Paz 64   15 Con actividad laboral o trabajo (0.2343750 0.7656250) *
##       15) Tarea7=1.Si 521  108 Con actividad laboral o trabajo (0.2072937 0.7927063) *
```

```
#Examinar los parámetros del árbol con printcp
printcp(mod1)
```

```
##
## Classification tree:
## rpart(formula = condac ~ ., data = trainbd)
##
## Variables actually used in tree construction:
## [1] area   depto  edad   Tarea7
##
## Root node error: 1308/5784 = 0.22614
##
## n= 5784
##
##         CP nsplit rel error  xerror     xstd
## 1 0.093272      0   1.00000 1.00000 0.024324
## 2 0.046636      2   0.81346 0.81651 0.022561
## 3 0.012997      3   0.76682 0.78670 0.022236
## 4 0.010000      5   0.74083 0.76988 0.022048
```

```
#Usar el comando plotcp para explorar los parámetros de forma gráfica
plotcp(mod1)
```




```
#Usar la función summary para para examinar el modelo
summary(mod1)
```

```
## Call:
## rpart(formula = condac ~ ., data = trainbd)
##   n= 5784
##
##           CP nsplit rel error    xerror       xstd
## 1 0.09327217      0 1.0000000 1.0000000 0.02432355
## 2 0.04663609      2 0.8134557 0.8165138 0.02256060
## 3 0.01299694      3 0.7668196 0.7866972 0.02223623
## 4 0.01000000      5 0.7408257 0.7698777 0.02204806
##
## Variable importance
##             area           Tarea7            depto             edad
##               44               23               21                4
##         inscrito         sabeLeer maxNivelAprobado maxCursoAprobado
##                3                2                2                1
##           Tarea5
##                1
##
## Node number 1: 5784 observations,    complexity param=0.09327217
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.2261411  P(node) =1
##     class counts:  4476  1308
##    probabilities: 0.774 0.226
##   left son=2 (4219 obs) right son=3 (1565 obs)
```

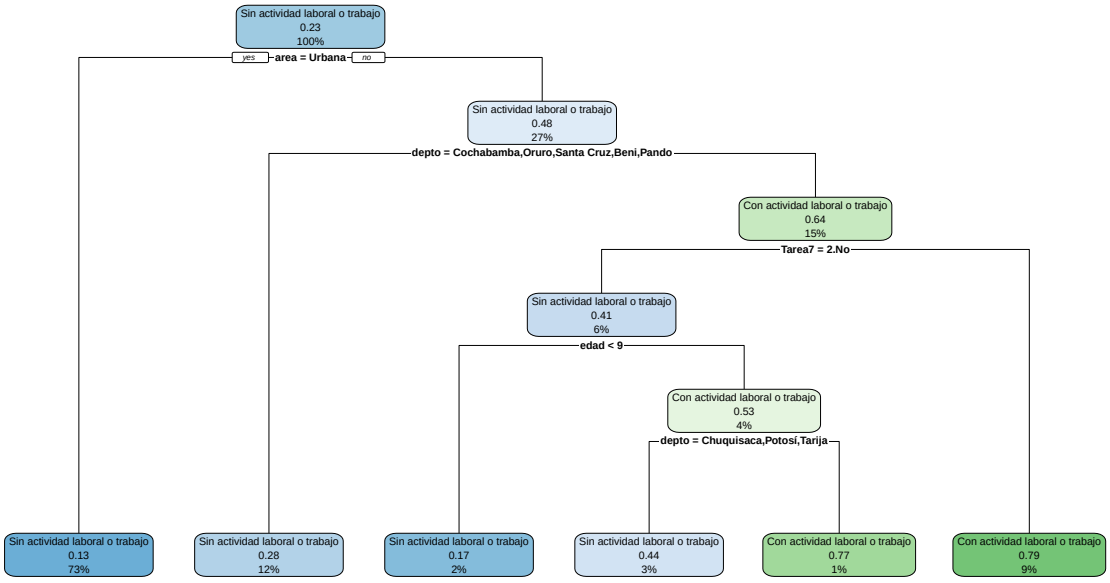
```
##   Primary splits:
##       area   splits as  LR,        improve=272.09680, (0 missing)
##       Tarea7 splits as  RL,        improve=203.30490, (0 missing)
##       depto  splits as  RLLLRLLLL, improve= 92.64829, (0 missing)
##       edad   < 10.5 to the left,   improve= 79.09017, (0 missing)
##       Tarea5 splits as  RL,        improve= 70.20063, (0 missing)
##   Surrogate splits:
##       Tarea7   splits as  RL,        agree=0.806, adj=0.282, (0 split)
##       depto    splits as  LLLLRLLLR, agree=0.743, adj=0.051, (0 split)
##       inscrito splits as  LLLR,      agree=0.743, adj=0.049, (0 split)
##
## Node number 2: 4219 observations
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.1327329  P(node) =0.729426
##     class counts:  3659   560
##    probabilities: 0.867 0.133
##
## Node number 3: 1565 observations,    complexity param=0.09327217
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.4779553  P(node) =0.270574
##     class counts:   817   748
##    probabilities: 0.522 0.478
##   left son=6 (713 obs) right son=7 (852 obs)
##   Primary splits:
##       depto            splits as  RRLLRRLLL, improve=102.11980, (0 missing)
##       Tarea7           splits as  RL, improve= 67.95451, (0 missing)
##       edad             < 8.5  to the left,  improve= 65.32596, (0 missing)
##       Tarea5           splits as  RL, improve= 41.74621, (0 missing)
##       maxNivelAprobado splits as  L-L--RRR-R--------, improve= 40.28798, (0 missing)
##   Surrogate splits:
##       inscrito         splits as  RLLR, agree=0.563, adj=0.041, (0 split)
##       Tarea7           splits as  RL, agree=0.559, adj=0.032, (0 split)
##       asistencia       splits as  RL, agree=0.548, adj=0.007, (0 split)
##       maxNivelAprobado splits as  R-R--RRR-L--------, agree=0.545, adj=0.001, (0 split)
##       usual            splits as  R-L, agree=0.545, adj=0.001, (0 split)
##
## Node number 6: 713 observations
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.2805049  P(node) =0.1232711
##     class counts:   513   200
##    probabilities: 0.719 0.281
##
## Node number 7: 852 observations,    complexity param=0.04663609
##   predicted class=Con actividad laboral o trabajo  expected loss=0.3568075  P(node) =0.1473029
##     class counts:   304   548
##    probabilities: 0.357 0.643
##   left son=14 (331 obs) right son=15 (521 obs)
##   Primary splits:
##       Tarea7           splits as  RL, improve=59.95731, (0 missing)
##       edad             < 7.5  to the left,  improve=57.13707, (0 missing)
##       sabeLeer         splits as  RL, improve=38.27188, (0 missing)
##       maxNivelAprobado splits as  L-L--RRR----------, improve=35.24853, (0 missing)
##       Tarea5           splits as  RL, improve=32.72622, (0 missing)
##   Surrogate splits:
##       depto            splits as  RR--RL---, agree=0.635, adj=0.060, (0 split)
##       edad             < 6.5  to the left,  agree=0.630, adj=0.048, (0 split)
##       inscrito         splits as  RL-R, agree=0.627, adj=0.039, (0 split)
```

```
##       sabeLeer         splits as  RL, agree=0.622, adj=0.027, (0 split)
##       maxNivelAprobado splits as  L-L--RRL----------, agree=0.620, adj=0.021, (0 split)
##
## Node number 14: 331 observations,    complexity param=0.01299694
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.407855  P(node) =0.05722683
##     class counts:   196   135
##    probabilities: 0.592 0.408
##   left son=28 (113 obs) right son=29 (218 obs)
##   Primary splits:
##       edad             < 8.5  to the left,  improve=19.718080, (0 missing)
##       Tarea5           splits as  RL, improve=15.752020, (0 missing)
##       maxNivelAprobado splits as  L-L--LRR----------, improve=14.216920, (0 missing)
##       sabeLeer         splits as  RL, improve=11.573830, (0 missing)
##       depto            splits as  LR--LL---, improve= 6.633786, (0 missing)
##   Surrogate splits:
##       maxCursoAprobado < 2.5  to the left,  agree=0.816, adj=0.460, (0 split)
##       sabeLeer         splits as  RL, agree=0.813, adj=0.451, (0 split)
##       maxNivelAprobado splits as  L-L--RRR----------, agree=0.813, adj=0.451, (0 split)
##       Tarea5           splits as  RL, agree=0.776, adj=0.345, (0 split)
##       asistencia       splits as  RL, agree=0.662, adj=0.009, (0 split)
##
## Node number 15: 521 observations
##   predicted class=Con actividad laboral o trabajo  expected loss=0.2072937  P(node) =0.09007607
##     class counts:   108   413
##    probabilities: 0.207 0.793
##
## Node number 28: 113 observations
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.1681416  P(node) =0.01953665
##     class counts:    94    19
##    probabilities: 0.832 0.168
##
## Node number 29: 218 observations,    complexity param=0.01299694
##   predicted class=Con actividad laboral o trabajo  expected loss=0.4678899  P(node) =0.03769018
##     class counts:   102   116
##    probabilities: 0.468 0.532
##   left son=58 (154 obs) right son=59 (64 obs)
##   Primary splits:
##       depto            splits as  LR--LL---, improve=9.880410, (0 missing)
##       Tarea5           splits as  RL, improve=6.210742, (0 missing)
##       edad             < 12.5 to the left,  improve=4.606581, (0 missing)
##       maxNivelAprobado splits as  -----LRR----------, improve=2.855229, (0 missing)
##       Tarea8           splits as  LR, improve=2.521765, (0 missing)
##   Surrogate splits:
##       maxNivelAprobado splits as  -----LLR----------, agree=0.711, adj=0.016, (0 split)
##
## Node number 58: 154 observations
##   predicted class=Sin actividad laboral o trabajo  expected loss=0.4350649  P(node) =0.02662517
##     class counts:    87    67
##    probabilities: 0.565 0.435
##
## Node number 59: 64 observations
##   predicted class=Con actividad laboral o trabajo  expected loss=0.234375  P(node) =0.01106501
##     class counts:    15    49
##    probabilities: 0.234 0.766
```

```
#Visualizar el árbol
#install.packages("rpart.plot")
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.2
```

```
rpart.plot(mod1)
```



Interpretación

```
clase<-predict(mod1,testbd,type = "class")
table(clase,testbd$condac)
```

```
##
## clase                              Sin actividad laboral o trabajo
##   Sin actividad laboral o trabajo                              1869
##   Con actividad laboral o trabajo                                66
##
## clase                              Con actividad laboral o trabajo
##   Sin actividad laboral o trabajo                               316
##   Con actividad laboral o trabajo                               190
```

```
#install.packages('caret')
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
confusionMatrix(table(clase,testbd$condac))
```

```
## Confusion Matrix and Statistics
##
##
## clase                              Sin actividad laboral o trabajo
##   Sin actividad laboral o trabajo                            1869
##   Con actividad laboral o trabajo                              66
##
## clase                              Con actividad laboral o trabajo
##   Sin actividad laboral o trabajo                             316
##   Con actividad laboral o trabajo                             190
##
##                Accuracy : 0.8435
##                  95% CI : (0.8285, 0.8577)
##     No Information Rate : 0.7927
##     P-Value [Acc > NIR] : 9.343e-11
##
##                   Kappa : 0.4176
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9659
##             Specificity : 0.3755
##          Pos Pred Value : 0.8554
##          Neg Pred Value : 0.7422
##              Prevalence : 0.7927
##          Detection Rate : 0.7657
##    Detection Prevalence : 0.8951
##       Balanced Accuracy : 0.6707
##
##        'Positive' Class : Sin actividad laboral o trabajo
##
```

6 Resultados y análisis

7 Conclusiones y recomendaciones

8 Referencias