

# Tarea3

Jason Solano and Percy Herrera

3/6/2020

## Tarea 3

En este documento se presenta un análisis parcial con la metodología CRISP-DM, en la cual se aplica un modelo descriptivo utilizando reglas de asociación. Por otra parte se interpretara dichas reglas para un análisis mas detallado.

## Librerías

### Librerías utilizadas

```
library(arules)
```

```
## Loading required package: Matrix
```

```
##  
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':  
##  
##      abbreviate, write
```

```
library(arulesViz)
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'seriation':  
##      method      from  
##      reorder.hclust gclus
```

```
library(tidyverse)
```

```
## — Attaching packages —  
—— tidyverse 1.3.0 —
```

```
## ✓ ggplot2 3.3.0      ✓ purrr 0.3.3
## ✓ tibble 2.1.3       ✓ dplyr 0.8.4
## ✓ tidyr 1.0.2        ✓ stringr 1.4.0
## ✓ readr 1.3.1        ✓ forcats 0.5.0
```

```
## — Conflicts ————— tidyverse_conflicts() —
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x tidyr::pack()    masks Matrix::pack()
## x dplyr::recode()  masks arules::recode()
## x tidyr::unpack() masks Matrix::unpack()
```

```
library(plyr)
```

```
## -----
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then d
plyr:
## library(plyr); library(dplyr)
```

```
## -----
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##      arrange, count, desc, failwith, id, mutate, rename, summarise,
##      summarize
```

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

```
library(ggplot2)
library(knitr)
library(lubridate)
```

```
##  
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:plyr':  
##  
##     here
```

```
## The following object is masked from 'package:base':  
##  
##     date
```

```
library(RColorBrewer)  
library(arules)
```

# Utilización de Datos

Cargamos el dataset que contiene información de residentes adultos de Estados Unidos

```
datUSA<- read.csv('AdultosUSA.csv',sep=';',dec=',',stringsAsFactors = FALSE)  
view(datUSA)
```

# Limpieza de datos

Removemos los datos nulos

```
datosCompletosUSA<- na.omit(datUSA)  
view(datosCompletosUSA)
```

Removemos datos con valores incongruentes

```
datosCompletosUSA <- datUSA[!(datUSA$Edad==" " | datUSA$TipoTrabajo==" " | datUSA$Ni  
velEducativo==" " | datUSA$NivelEducativo==" " | datUSA$AnosEducacion==" " | datUSA$Estad  
oCivil==" " | datUSA$Ocupacion==" " | datUSA$Sexo==" " | datUSA$HorasSemanales==" " | datUSA$  
PaisOrigen==" " | datUSA$Ingresos==" " ), ]  
view(datosCompletosUSA)
```

Observamos los datos

```
summary(datosCompletosUSA)
```

```
##      Edad      TipoTrabajo      NivelEducativo      AnnosEducacion
## Min.      :17.00    Length:25000    Length:25000    Min.      : 1.00
## 1st Qu.:28.00    Class :character    Class :character    1st Qu.: 9.00
## Median :37.00    Mode  :character    Mode  :character    Median :10.00
## Mean   :38.61
## 3rd Qu.:48.00
## Max.    :90.00
##
## EstadoCivil      Ocupacion      Sexo      HorasSemanales
## Length:25000    Length:25000    Length:25000    Min.      : 1.00
## Class :character    Class :character    Class :character    1st Qu.:40.00
## Mode  :character    Mode  :character    Mode  :character    Median :40.00
##
## Mean   :40.41
## 3rd Qu.:45.00
## Max.    :99.00
##
## PaisOrigen      Ingresos
## Length:25000    Length:25000
## Class :character    Class :character
## Mode  :character    Mode  :character
##
##
##
```

## Inicamos con la transformación de

Las variables son no numéricas Se convierten a factor las variables TipoTrabajo, NivelEducativo, EstadoCivil, Ocupacion, Sexo, PaisOrigen y Ingresos

```
datosCompletosUSA$TipoTrabajo<- as.factor(datosCompletosUSA$TipoTrabajo)

datosCompletosUSA$NivelEducativo<- as.factor(datosCompletosUSA$NivelEducativo)

datosCompletosUSA$EstadoCivil<- as.factor(datosCompletosUSA$EstadoCivil)

datosCompletosUSA$Ocupacion<- as.factor(datosCompletosUSA$Ocupacion)

datosCompletosUSA$Sexo<- as.factor(datosCompletosUSA$Sexo)

datosCompletosUSA$PaisOrigen<- as.factor(datosCompletosUSA$PaisOrigen)

datosCompletosUSA$Ingresos<- as.factor(datosCompletosUSA$Ingresos)
```

## Definición de Rangos

Para el caso de horas semanales, se definen los siguientes rangos, debido a que 20 horas son trabajos de medio tiempo y 40 horas son trabajos de tiempo completo

```
datosCompletosUSA$HorasSemanales <- discretize(datosCompletosUSA$HorasSemanales,
method="fixed",breaks = c(0, 20, 40,60,100))
```

Para el caso de edad, se empieza los rangos de 0 a 20 donde hay pocos ingresos, de 20 a 40 donde existe un gran crecimiento de ingresos y de 40 a 60 cerca de la pensión, y de 60 a 100 donde existen gran cantidad de salarios con pensión

```
datosCompletosUSA$Edad <- discretize(datosCompletosUSA$Edad, method="fixed",break
s = c(0, 20, 40,60,100))
```

Para el caso de educación se divide de 0 a 5 años para el caso de personas con educación en primaria, secundaria de 5 a 10, para el caso universitario tipo bachillerato de 10 a 15, y maestrías o doctorados de 15 a 30

```
datosCompletoUSA$AnosEducacion <- discretize(datosCompletoUSA$AnosEducacion,
method="fixed",breaks = c(0, 5, 10,15,30))
```

## Observamos los datos

```
summary(datosCompletoUSA)
```

##	Edad	TipoTrabajo	NivelEducativo	AnnosEducacion
##	[0,20) : 1257	Private :17385	HS-grad :8120	[0,5) : 891
##	[20,40) :12793	Self-emp-not-inc: 1978	Some-college:5597	[5,10) :10467
##	[40,60) : 8930	Local-gov : 1624	Bachelors :4140	[10,15):12897
##	[60,100]: 2020	? : 1399	Masters :1300	[15,30]: 745
##		State-gov : 993	Assoc-voc :1059	
##		Self-emp-inc : 857	11th : 909	
##		(Other) : 764	(Other) :3875	
##	EstadoCivil	Ocupacion	Sexo	
##	Divorced : 3435	Prof-specialty :3180	Female: 8291	
##	Married-AF-spouse : 16	Craft-repair :3122	Male :16709	
##	Married-civ-spouse :11441	Exec-managerial:3084		
##	Married-spouse-absent: 328	Adm-clerical :2975		
##	Never-married : 8225	Sales :2815		
##	Separated : 786	Other-service :2555		
##	Widowed : 769	(Other) :7269		
##	HorasSemanales	PaisOrigen	Ingresos	
##	[0,20) : 1301	United-States:22421	<=50K.:19016	
##	[20,40) : 4673	Mexico : 488	>50K. : 5984	
##	[40,60) :17034	? : 445		
##	[60,100]: 1992	Philippines : 151		
##		Germany : 102		
##		Canada : 99		
##		(Other) : 1294		

# Generación de reglas

Primero para generar reglas debemos convertir nuestros datos a tipo basket, para lo cual utilizamos el siguiente comando

```
trns<- as(datosCompletoUSA,'transactions')

summary(trns)
```

```
## transactions as itemMatrix in sparse format with
## 25000 rows (elements/itemsets/transactions) and
## 105 columns (items) and a density of 0.0952381
##
## most frequent items:
## PaisOrigen=United-States      Ingresos=<=50K.      TipoTrabajo=Private
##                22421                19016                17385
##   HorasSemanales=[40,60)      Sexo=Male      (Other)
##                17034                16709                157435
##
## element (itemset/transaction) length distribution:
## sizes
##    10
## 25000
##
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10      10      10      10      10      10
##
## includes extended item information - examples:
##      labels variables  levels
## 1  Edad=[0,20)      Edad  [0,20)
## 2  Edad=[20,40)     Edad  [20,40)
## 3  Edad=[40,60)     Edad  [40,60)
##
## includes extended transaction information - examples:
##      transactionID
## 1                1
## 2                2
## 3                3
```

Se generan las reglas relacionadas con ingresos superiores a 50k:

```
reglasIngresos50p<- apriori(trns,parameter =list(supp=0.001,conf=0.8),
                             appearance = list(default='lhs',rhs='Ingresos=>50K. ')
)
```

```
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##          0.8    0.1    1 none FALSE                TRUE         5   0.001      1
## maxlen target   ext
##       10  rules FALSE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##    0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 25
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[105 item(s), 25000 transaction(s)] done [0.03s].
## sorting and recoding items ... [85 item(s)] done [0.00s].
## creating transaction tree ... done [0.02s].
## checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.21s].
## writing ... [1956 rule(s)] done [0.02s].
## creating S4 object ... done [0.02s].
```

Observamos la cantidad de reglas

```
length(reglasIngresos50p)
```

```
## [1] 1956
```

Filtramos las primeras reglas

```
inspect(head(reglasIngresos50p))
```

##	lhs	rhs	support	confidence
	lift count			
## [1]	{TipoTrabajo=Self-emp-inc, NivelEducativo=Doctorate}	=> {Ingresos=>50K.}	0.00100	0.8333333 3.
481506	25			
## [2]	{TipoTrabajo=State-gov, NivelEducativo=Doctorate}	=> {Ingresos=>50K.}	0.00248	0.8378378 3.
500325	62			
## [3]	{NivelEducativo=Doctorate, Ocupacion=Exec-managerial}	=> {Ingresos=>50K.}	0.00160	0.8888889 3.
713607	40			
## [4]	{Edad=[40,60), NivelEducativo=Doctorate}	=> {Ingresos=>50K.}	0.00596	0.8232044 3.
439190	149			
## [5]	{NivelEducativo=Doctorate, EstadoCivil=Married-civ-spouse}	=> {Ingresos=>50K.}	0.00708	0.8309859 3.
471699	177			
## [6]	{TipoTrabajo=Self-emp-inc, NivelEducativo=Prof-school}	=> {Ingresos=>50K.}	0.00212	0.9464286 3.
953996	53			

Parte de la limpieza de reglas consta con eliminar la reglas que son subconjunto de otras

```
subconjuntos<- which(colSums(is.subset(reglasIngresos50p,reglasIngresos50p))>1)

reglasFinal<- reglasIngresos50p[-subconjuntos]
```

Observamos la cantidad de reglas

```
length(reglasFinal)
```

```
## [1] 101
```

Observamos las mejores 10 reglas

```
inspect(sort(reglasFinal,by='support',decreasing = TRUE)[1:10])
```

##	lhs	rhs	support	confidence
	lift count			
## [1]	{Edad=[40,60), AnnosEducacion=[10,15), EstadoCivil=Married-civ-spouse, Ocupacion=Exec-managerial}	=> {Ingresos=>50K.}	0.02384	0.8000000 3
.342246	596			
## [2]	{Edad=[40,60),			



```

##      TipoTrabajo=Private,
##      EstadoCivil=Married-civ-spouse,
##      Ocupacion=Exec-managerial}      => {Ingresos=>50K.} 0.01844 0.8045375 3
.361203 461
## [3] {AnnosEducacion=[15,30],
##      EstadoCivil=Married-civ-spouse} => {Ingresos=>50K.} 0.01744 0.8304762 3
.469570 436
## [4] {AnnosEducacion=[15,30],
##      Ocupacion=Prof-specialty,
##      Sexo=Male}                      => {Ingresos=>50K.} 0.01568 0.8016360 3
.349081 392
## [5] {Edad=[40,60),
##      NivelEducativo=Masters,
##      EstadoCivil=Married-civ-spouse} => {Ingresos=>50K.} 0.01560 0.8315565 3
.474083 390
## [6] {NivelEducativo=Bachelors,
##      EstadoCivil=Married-civ-spouse,
##      Ocupacion=Exec-managerial,
##      Sexo=Male,
##      HorasSemanales=[40,60)}          => {Ingresos=>50K.} 0.01528 0.8093220 3
.381192 382
## [7] {Edad=[40,60),
##      TipoTrabajo=Private,
##      AnnosEducacion=[10,15),
##      Ocupacion=Exec-managerial,
##      Sexo=Male,
##      PaisOrigen=United-States}        => {Ingresos=>50K.} 0.01492 0.8021505 3
.351231 373
## [8] {NivelEducativo=Bachelors,
##      EstadoCivil=Married-civ-spouse,
##      Ocupacion=Exec-managerial,
##      HorasSemanales=[40,60),
##      PaisOrigen=United-States}        => {Ingresos=>50K.} 0.01468 0.8065934 3
.369792 367
## [9] {TipoTrabajo=Private,
##      NivelEducativo=Bachelors,
##      EstadoCivil=Married-civ-spouse,
##      Ocupacion=Exec-managerial}      => {Ingresos=>50K.} 0.01396 0.8060046 3
.367332 349
## [10] {Edad=[40,60),
##      AnnosEducacion=[15,30]}          => {Ingresos=>50K.} 0.01316 0.8225000 3
.436247 329

```

Observamos las tres mejores

```

tresMejores <- head(sort(reglasFinal,by='support',decreasing = TRUE)[1:3])
inspect(tresMejores)

```

##	lhs	rhs	support	confidence
	lift count			
## [1]	{Edad=[40,60),			
##	AnnosEducacion=[10,15),			
##	EstadoCivil=Married-civ-spouse,			
##	Ocupacion=Exec-managerial}	=> {Ingresos=>50K.}	0.02384	0.8000000 3.
342246	596			
## [2]	{Edad=[40,60),			
##	TipoTrabajo=Private,			
##	EstadoCivil=Married-civ-spouse,			
##	Ocupacion=Exec-managerial}	=> {Ingresos=>50K.}	0.01844	0.8045375 3.
361203	461			
## [3]	{AnnosEducacion=[15,30],			
##	EstadoCivil=Married-civ-spouse}	=> {Ingresos=>50K.}	0.01744	0.8304762 3.
469570	436			

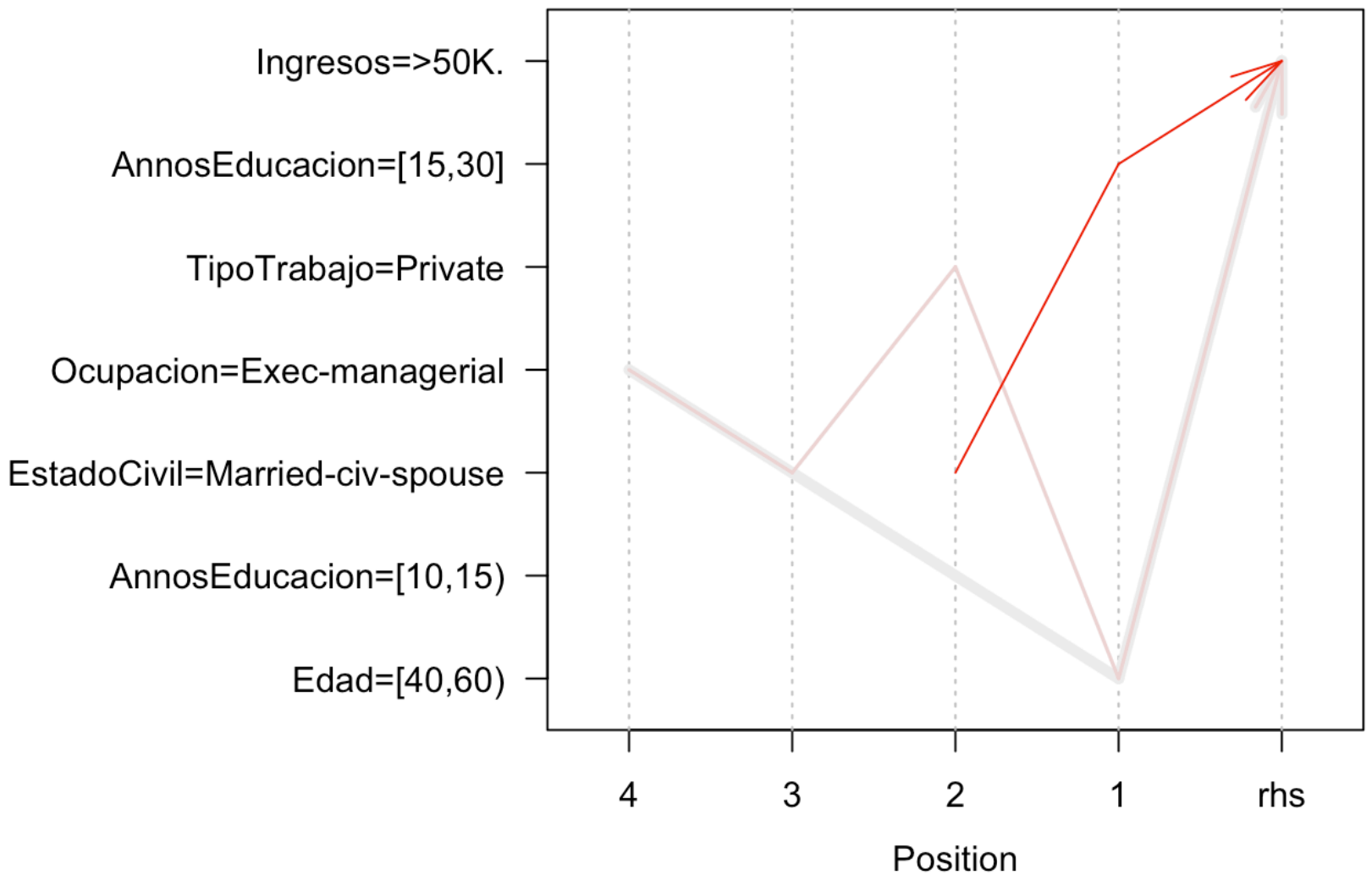
# Se gráfica el resultado

```
plot(tresMejores,method = 'graph',engine = 'htmlwidget')
```

Select by id

```
plot(tresMejores,method ='paracoord' )
```

## Parallel coordinates plot for 3 rules



## Conclusión

Como el anunciado de la tarea se indicaba el objetivo del negocio y de la mineria de datos es encontrar patrones y reglas que permitan identificar cuando una persona gana mas de 50 mil dolares para este caso se concluye lo siguiente

Patrones:

- Una persona con un grado de educación unversitario se puede deducir tipo maestria o doctorado, universitario tipo bachillerato
- Se su trabajo en el sector privado
- Edad Persona mayor de 40 años y menor de 60 años
- Ocupación gerencial
- Estado civil: casado de forma legal

Reglas:

- Edad=[40,60),AnnosEducacion=[10,15),stadoCivil=Married-civ-spouse,Ocupacion=Exec-managerial
- Edad=[40,60) TipoTrabajo=Private, EstadoCivil=Married-civ-spouse, Ocupacion=Exec-managerial

- AnnosEducacion=[15,30] EstadoCivil=Married-civ-spouse

Por lo tanto el negocio debe prestar atención a los clientes con dichas características para ofrecer sus productos premium o de mayor costo que puedan ser comprados por personas que ganen mas de 50 mil dolares al año.