

MINE 4204 - Laboratorio - Control estadístico de divulgación de datos

Departamento de Ingeniería de Sistemas y Computación

Curso: MINE 4204 Información, Seguridad y Privacidad

Semestre 2018-2

- [MINE 4204 - Laboratorio - Control estadístico de divulgación de datos](#)
 - [Autores](#)
 - [Respuestas](#)
 - [Sección 1](#)
 - [a. ¿Hay relaciones entre algunos de ellos?](#)
 - [b. ¿Es necesario suprimir algún atributo?](#)
 - [c. ¿Cuáles son deducibles a partir de otros?](#)
 - [d. ¿Cuáles son categóricos y cuáles numéricos?](#)
 - [Sección 2](#)
 - [a. Indique los cambios realizados.](#)
 - [Sección 3](#)
 - [Sección 4](#)
 - [a. ¿Hay observaciones únicas? ¿Cuántas?](#)
 - [b. ¿El riesgo de re-identificación es muy alto?](#)
 - [c. Copie los comandos ejecutados para cumplir con este punto.](#)
 - [Sección 5](#)
 - [a. Copie los comandos ejecutados para cumplir con este punto.](#)
 - [Sección 6](#)
 - [a. Copie los comandos ejecutados para cumplir con este punto.](#)

Autores

Marly Piedrahita ([mj.piedrahita](#))

Juan Méndez ([jc.mendez](#))

Source Code:

https://github.com/dersteppenwolf/mine4204/tree/master/lab_divulgaciondatos

Respuestas

Sección 1

a. ¿Hay relaciones entre algunos de ellos?

Para determinar si hay relaciones entre atributos se realizó un análisis de correlación.

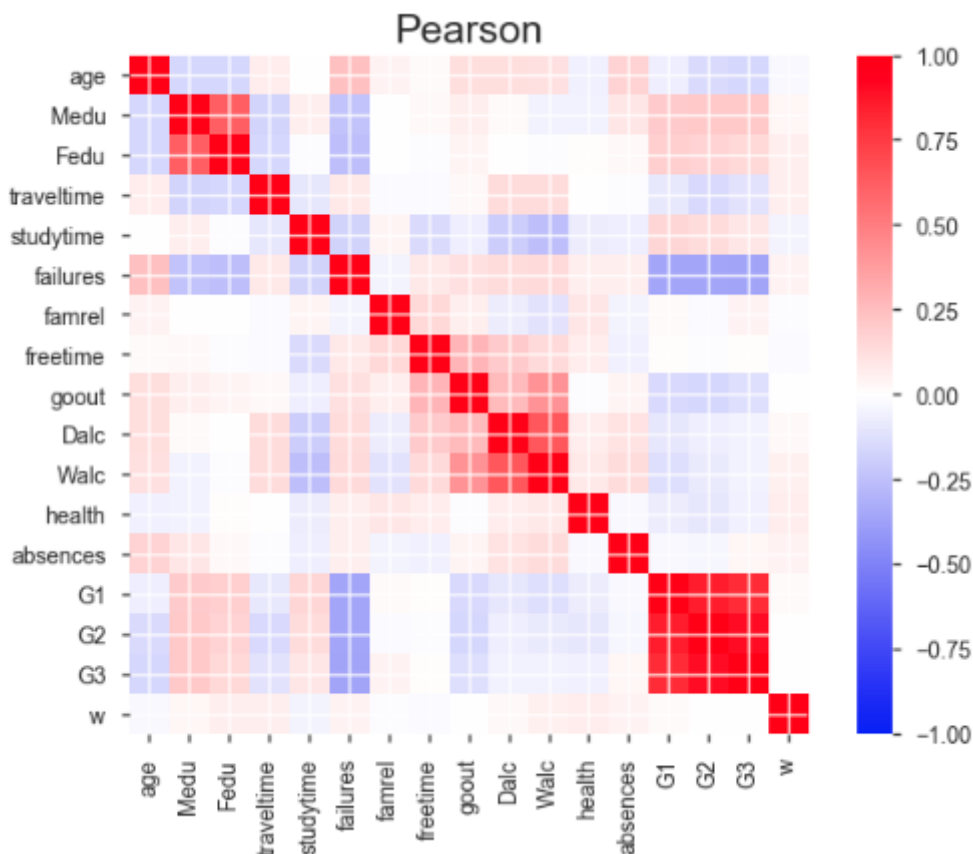
Para la generación de dicho análisis, además de la revisión descriptiva de los datos, se utilizó la herramienta Pandas Profiling (<https://github.com/pandas-profiling/pandas-profiling>). (Nota: Puede encontrar el notebook de Jupyter en el archivo **data_quality.ipynb** que se encuentra en esta misma carpeta)

Se encontró correlación entre algunas variables tales como:

Medu / Fedu

Dalc / Walc

G1 / G2 / G3



b. ¿Es necesario suprimir algún atributo?

No. Dado que en el dataset hay datos que sirvan como identificadores únicos de los individuos.

c. ¿Cuáles son deducibles a partir de otros?

Para los atributos que tienen una correlación alta es probable que se puedan hacer deducciones a través de otros.

Para el caso de este dataset tenemos que hay correlación entre los siguientes atributos:

Medu / Fedu
Dalc / Walc
G1 / G2 / G3

d. ¿Cuáles son categóricos y cuáles numéricos?

La clasificación de Categóricos y Numéricos se obtuvo a partir de la operación `info()` del dataframe de pandas (<https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.info.html>)

Categoricas: 28

school	395 non-null object
sex	395 non-null object
address	395 non-null object
famsize	395 non-null object
Pstatus	395 non-null object
Mjob	395 non-null object
Fjob	395 non-null object
reason	395 non-null object
guardian	395 non-null object
famsup	395 non-null object
paid	395 non-null object
activities	395 non-null object
nursery	395 non-null object
higher	395 non-null object
internet	395 non-null object
romantic	395 non-null object
schoolsup	395 non-null object
Medu	395 non-null int64
Fedu	395 non-null int64
traveltime	395 non-null int64
studytime	395 non-null int64
failures	395 non-null int64
famrel	395 non-null int64
freetime	395 non-null int64
goout	395 non-null int64
Dalc	395 non-null int64
Walc	395 non-null int64
health	395 non-null int64

Atributos numéricos 5:

age	395 non-null int64
absences	395 non-null int64
G1	395 non-null int64
G2	395 non-null int64
G3	395 non-null int64
w	395 non-null int64

Sección 2

Para los atributos binarios y nominales que así lo requieran, cambie las etiquetas de texto por identificadores numéricos (e.g. en los binarios, cambie "yes" por 1 y "no" por 0).

a. Indique los cambios realizados.

Procedimiento Realizado:

Para la conversión de las variables categóricas en numéricas se realizó el siguiente procedimiento:

1. Convertir los atributos de tipo object en variable categórica:

```
df['school'] = df['school'].astype('category')
df['sex'] = df['sex'].astype('category')
df['address'] = df['address'].astype('category')
df['famsize'] = df['famsize'].astype('category')
df['Pstatus'] = df['Pstatus'].astype('category')
df['Mjob'] = df['Mjob'].astype('category')
df['Fjob'] = df['Fjob'].astype('category')
df['reason'] = df['reason'].astype('category')
df['guardian'] = df['guardian'].astype('category')
df['famsup'] = df['famsup'].astype('category')
df['paid'] = df['paid'].astype('category')
df['activities'] = df['activities'].astype('category')
df['nursery'] = df['nursery'].astype('category')
df['higher'] = df['higher'].astype('category')
df['internet'] = df['internet'].astype('category')
df['romantic'] = df['romantic'].astype('category')
df['schoolsup'] = df['schoolsup'].astype('category')
```

2. Codificar las variables categóricas como numéricas

```
cat_columns = df.select_dtypes(['category']).columns
df[cat_columns] = df[cat_columns].apply(lambda x: x.cat.codes)
```

3. Exportar los datos resultantes como csv (**dataset_tarea.csv**)

```
df.to_csv('dataset_tarea.csv')
```

Ejemplo de la tabla recodificada:

Sample

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	f
0	0	0	18	1	0	0	4	4	0	4	0	1	2	2	0	1	
1	0	0	17	1	0	1	1	1	0	2	0	0	1	2	0	0	
2	0	0	15	1	1	1	1	1	0	2	2	1	1	2	3	1	
3	0	0	15	1	0	1	4	2	1	3	1	1	1	3	0	0	
4	0	0	16	1	0	1	3	3	2	2	1	0	1	2	0	0	

Sección 3

Cargue el archivo resultante a RStudio. Para ello, puede utilizar el siguiente comando:

```
<suFrame> =read.table("C:/la_ruta_a_su_archivo.csv", sep=";", header=TRUE)
```

Comando Utilizado

```
library(sdcMicro)
```

```
df2 <- read.table("/lab_divulgaciondatos/dataset_tarea.csv", header=TRUE, sep=";
```

```
nrow(df2)
```

```
ncol(df2)
```

```
print(df2)
```

Ejemplo Salida R Studio:

```

> df2 <- read.table("/Users/ivanmatis/Downloads/University/security/mine4204/lab_divulgaciondatos/dataset_tarea.csv", header=TRUE, sep = ",")
>
> nrow(df2)
[1] 395
> ncol(df2)
[1] 35
>
> print(df2)

```

	X	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
1	0	0	0	18	1	0	0	4	4	0	4	0	1
2	1	0	0	17	1	0	1	1	1	0	2	0	0
3	2	0	0	15	1	1	1	1	1	0	2	2	1
4	3	0	0	15	1	0	1	4	2	1	3	1	1
5	4	0	0	16	1	0	1	3	3	2	2	1	0
6	5	0	1	16	1	1	1	4	3	3	2	3	1

Sección 4

Puesto que ya identificó qué variables son categóricas, puede proceder con la anonimización de dichas variables. Empiece identificando las frecuencias y riesgos de divulgación. Una vez tenga conocimiento de la existencia de combinaciones únicas de variables, utilice los métodos de recodificación global y supresión local para reducir el riesgo asociado. Puede graficar este último y valerse de los métodos previamente presentados.

a. ¿Hay observaciones únicas? ¿Cuántas?

Todas las 395 son únicas. observaciones se identificaron porque violan 2-anonymity

```

ff <- freqCalc(df2, keyVars=c('school', 'sex', 'address', 'famsize', 'Pstatus',
                             'reason', 'guardian', 'schoolsup', 'famsup', 'pa
                             'nursery', 'higher', 'internet', 'romantic',
                             "Medu", "Fedu", "traveltime",
                             "studytime", "failures", "famrel", "freetime",
                             "goout", "Dalc", "Walc", "health"), w=35)

print(cbind(df2, ff$fk, ff$Fk))

print(ff)
#-----
# 395 obs. violate 2-anonymity
# 395 obs. violate 3-anonymity
#-----

```

b. ¿El riesgo de re-identificación es muy alto?

El riesgo es muy alto ya que el 100% de los registros del dataset son posible identificarlos de forma única a través de la combinación de sus variables categóricas.

```
#calculo del riesgo
```

```
rk <- indivRisk(ff)$rk
print (cbind(df2, ff$fk, ff$Fk, rk ))
```

c. Copie los comandos ejecutados para cumplir con este punto.

Nota: Puede suponer que se cumple el supuesto de distribución Poisson de *fk* y *Fk*, porque los datos fueron especialmente preparados para ello.

Para la selección de las variables a tener en cuenta durante la recodificación se utilizan los atributos *mjob*, *fjob*, *reason* y *guardian* debido a que presentan más de dos valores donde uno de ellos tiene una frecuencia muy baja de fácil identificación:

Mjob

Numeric

Distinct count

Unique (%)

Missing (%)

Missing (n)

Infinite (%)

Infinite (n)

5

1.3%

0.0%

0

0.0%

0

Mean

Minimum

Maximum


Zeros (%)

2.1696

0

4

14.9%



Fjob

Numeric

Distinct count

Unique (%)

Missing (%)

Missing (n)

Infinite (%)

Infinite (n)

5

1.3%

0.0%

0

0.0%

0

Mean

Minimum

Maximum


Zeros (%)

2.281

0

4

5.1%



reason

Numeric

Distinct count

Unique (%)

Missing (%)

Missing (n)

Infinite (%)

Infinite (n)

4

1.0%

0.0%

0

0.0%

0

Mean

Minimum

Maximum


Zeros (%)

1.2557

0

3

36.7%



guardian

Numeric

Distinct count

Unique (%)

Missing (%)

Missing (n)

Infinite (%)

Infinite (n)

3

0.8%

0.0%

0

0.0%

0

Mean

Minimum

Maximum


Zeros (%)

0.85316

0

2

22.8%



Comandos para recodificar:

```
#Anonimización variables cartegóricas
```

```
#mjob
df2[,10] <- globalRecode(df2[,10] , breaks=c( -1, 2, 4 ) , labels=c( 1,2 ))
```

```
#fjob
df2[,11] <- globalRecode(df2[,11] , breaks=c( -1, 2, 4 ) , labels=c( 1,2 ))
```

```
#reason
df2[,12] <- globalRecode(df2[,12] , breaks=c( -1, 1, 3 ) , labels=c( 1,2 ))
```

```
#guardian
```

```
df2[,13] <- globalRecode(df2[,13] , breaks=c( -1, 0, 2 ) , labels=c( 1,2 ))

# calcular frecuencias de las tipo category después de anonimización

ff <- freqCalc(df2, keyVars=c('school', 'sex', 'address', 'famsize', 'Pstatus'
                              'reason', 'guardian', 'schoolsup', 'famsup', 'pa
                              'nursery', 'higher', 'internet', 'romantic',
                              "Medu", "Fedu", "traveltime",
                              "studytime", "failures", "famrel", "freetime",
                              "goout", "Dalc" , "Walc" , "health") , w=35)

print(cbind(df2,ff$fk,ff$Fk))

#calculo del riesgo después de anonimización

rk <- indivRisk(ff)$rk
print (cbind(df2, ff$fk, ff$Fk, rk ))
```

Luego de la recodificación se aplica el proceso de supresión de datos en todas los atributos categóricos para lograr la k anonymity de 2:

```
# Local Suppression To Obtain K-Anonymity
localsupx <- kAnon(df2, keyVars=c('school', 'sex', 'address', 'famsize', 'Psta
                              'reason', 'guardian', 'schoolsup', 'famsup',
                              'nursery', 'higher', 'internet', 'romantic',
                              "Medu", "Fedu", "traveltime",
                              "studytime", "failures", "famrel", "freetime
                              "goout", "Dalc" , "Walc" , "health"), k=2)

plot(localsupx)
print(localsupx$xAnon)

newX <- cbind(localsupx$xAnon, df2$w);
print(newX)
print ( grep("w", colnames(newX)) )
newff <- freqCalc(newX, keyVars=c('school', 'sex', 'address', 'famsize', 'Psta
                              'reason', 'guardian', 'schoolsup', 'famsup',
                              'nursery', 'higher', 'internet', 'romantic',
                              "Medu", "Fedu", "traveltime",
                              "studytime", "failures", "famrel", "freetime
                              "goout", "Dalc" , "Walc" , "health") , w=29)

print(cbind(newX,newff$fk,newff$Fk))

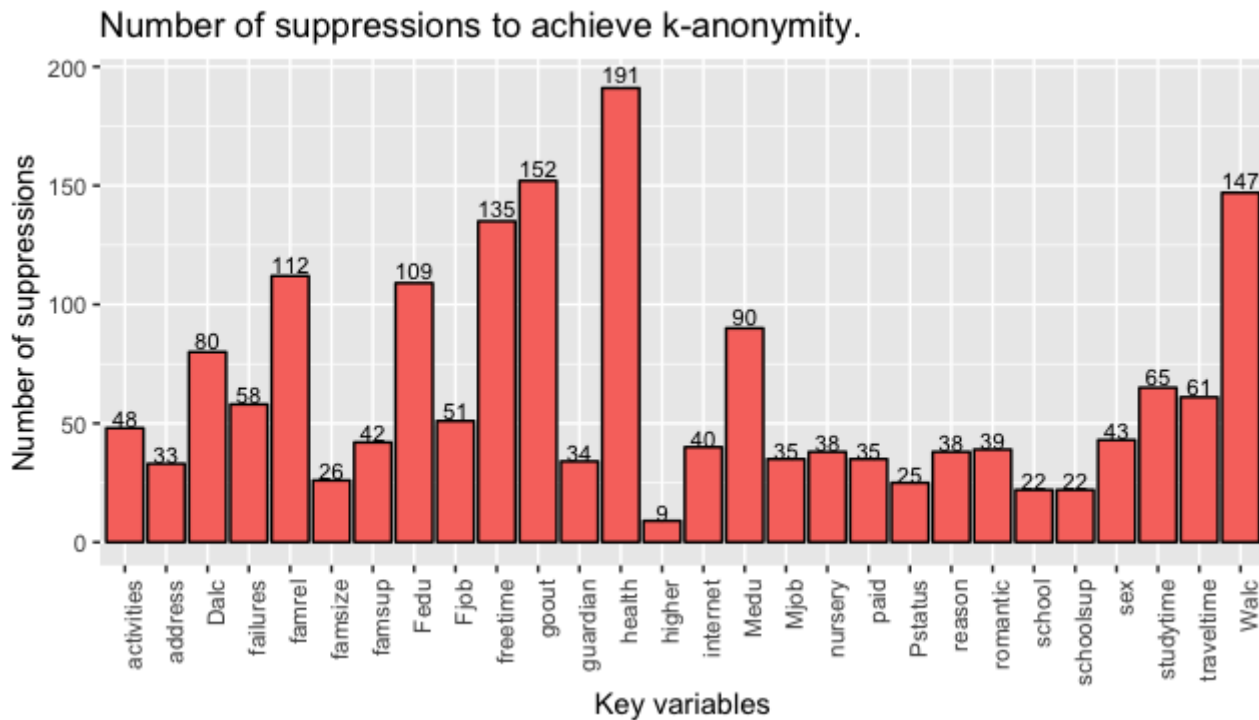
print(newff)
#-----
# 0 obs. violate 2-anonymity
# 252 obs. violate 3-anonymity
#-----
```



```
newrk <- indivRisk( newff )$rk
print (cbind(newX, newff$fk , newff$Fk , newrk ))
print(newrk)
```

Como resultado del proceso se obtiene que ningún registro viola la regla de k anonymity = 2, y que 252 registros violan la k anonymity de 3.

En el siguiente gráfico se encuentran enumeradas la cantidad de supresiones por atributo:



Sección 5

1. Es necesario continuar el proceso con las variables numéricas. Considere las posibilidades de añadir ruido, o efectuar rank swapping sobre alguno(s) de ellos. Posteriormente, utilice el procedimiento de microagregación para perturbar el conjunto de datos. Por último, no olvide evaluar estas transformaciones verificando la pérdida de información y el riesgo de divulgación. Note que estos dos últimos representan un trade-off a nivel del proceso de anonimización, porque el riesgo se disminuye sustancialmente a medida que aumenta el nivel de pérdida de información, pero, como es natural, no es deseable que este nivel crezca exponencialmente; razón por la cual es necesario encontrar un equilibrio entre ambos.

a. Copie los comandos ejecutados para cumplir con este punto.

```
#####
#Anonimización variables numéricas
```

```

# rankSwap
numDf <- rankSwap(df2, variables=c( "age" , "absences",
                                   "G1","G2", "G3" ),
                  TopPercent=20, BottomPercent=20, P = 15)

print(numDf)

numericalOnlyDF <- numDf[,c("age" , "absences",
                           "G1","G2", "G3")]
print(numericalOnlyDF)

rNoise <- addNoise(numericalOnlyDF,method="additive")
print(rNoise$xm)

nivelPerdidaNoise <- dUtility(numericalOnlyDF, xm=rNoise$xm, method="IL1")
print(nivelPerdidaNoise)

rMagg <- microaggregation(numericalOnlyDF, method ="mdav",aggr = 3)
print(rMagg$mx)

nivelPerdidaNum <- dUtility(numericalOnlyDF, xm=rMagg$xm, method="IL1")
print(nivelPerdidaNum)

print ( dRisk(numericalOnlyDF, rMagg$xm) )
print ( dRisk(df4, df4Resultadomdav$mx) )

```

Sección 6

Una vez haya verificado que el riesgo ha sido efectivamente reducido en ambos casos(variables categóricas y numéricas), unifique todo en un único dataset anonimizado que usted divulgaría sin temor de re- identificación de los estudiantes portugueses del estudio.

a. Copie los comandos ejecutados para cumplir con este punto.

Con los datos categóricos y numéricos procesados se genera el archivo dataset_anonimizado_tarea.csv :

```

## archivo final
# categoricas
print(newX)
# numericas
print(rMagg$mx)
#print(rMagg)

finalDf <- newX
print(finalDf)
catAndNum <- cbind(finalDf, rMagg$mx)

```

```
print(catAndNum)
```

```
write.csv(catAndNum, file = "dataset_anonimizado_tarea.csv")
```