Problem Set 1 RMarkdown

Irina Vélez - Lucía Fillippo - Daniel Casas - Miguel Victoria

2023-06-21

# Taller 1

El presente informe presenta la solución al Problem Set 1 de la clase Big Data & Machine Learning, en donde se aplicaron diversos conceptos y herramientas para la predicción de modelos, el manejo de bases de datos grandes, entre otros. El repositorio GitHub se encuentra este documento y el código R, en donde se generaron todos los resultados. Este repositorio se encuentra en el siguiente link: (link del Github con el código)

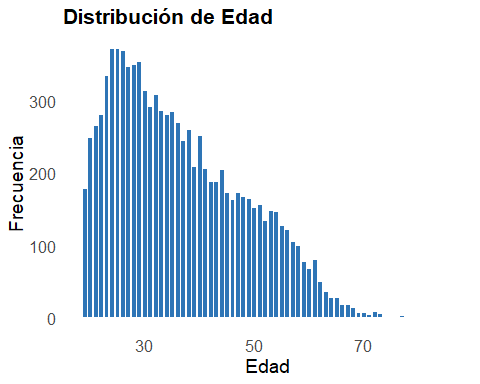
## 1. Introducción

Conocer el valor de los ingresos de las personas es fundamental para el pago de impuestos y para identificar a los hogares que requieren apoyos sociales; no obstante, en algunas ocasiones estos valores no se conocen al no ser reportados, de manera que poder determinar su valor se convierte en un gran insumo para el desarrollo de políticas tributarias y sociales. Con objeto de lo anterior, el objetivo principal de este documento es construir un modelo predictivo de los salarios por hora de los individuos, a partir del siguiente modelo:

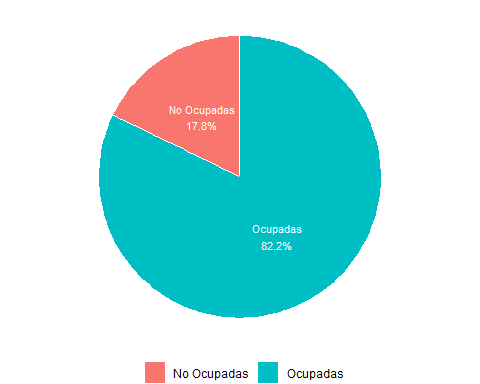
donde w representa el salario por hora y X es una matriz de potenciales variables que explican el salario. Para la creación de este modelo se utilizarán datos de la Gran Encuesta Integrada de Hogares – GEIH del año 2018.

Para importar los datos, es importante conocer qué tipo de página web contiene la información, en este caso, la página web que contiene las bases de datos es dinámica, por lo que es pertinente identificar el link principal del cual se realizará la extracción de la información; además, se apliará un código en bucle para que la extracción de la informacíon sea más eficiente.Por último, como fue indicado, se realiza un primer filtro para tener en cuenta únicamente los individuos mayores de 18 años.

## y\_salary\_m\_hu pet mes age   
## Min. : 151.9 Min. :1 Min. : 1.00 Min. :19.00   
## 1st Qu.: 3797.7 1st Qu.:1 1st Qu.: 4.00 1st Qu.:27.00   
## Median : 4522.3 Median :1 Median : 6.00 Median :34.00   
## Mean : 7984.7 Mean :1 Mean : 6.44 Mean :36.44   
## 3rd Qu.: 7291.7 3rd Qu.:1 3rd Qu.: 9.00 3rd Qu.:45.00   
## Max. :291666.7 Max. :1 Max. :12.00 Max. :86.00   
## sex ocu oficio maxEducLevel totalHoursWorked  
## Min. :0.0000 Min. :1 Min. : 1.00 Min. :1.000 Min. : 1.0   
## 1st Qu.:0.0000 1st Qu.:1 1st Qu.:33.00 1st Qu.:6.000 1st Qu.: 48.0   
## Median :1.0000 Median :1 Median :45.00 Median :6.000 Median : 48.0   
## Mean :0.5017 Mean :1 Mean :48.27 Mean :6.098 Mean : 48.4   
## 3rd Qu.:1.0000 3rd Qu.:1 3rd Qu.:58.00 3rd Qu.:7.000 3rd Qu.: 48.0   
## Max. :1.0000 Max. :1 Max. :99.00 Max. :7.000 Max. :130.0   
## exp   
## Min. : 0.000   
## 1st Qu.: 0.000   
## Median : 2.000   
## Mean : 3.989   
## 3rd Qu.: 5.000   
## Max. :58.000

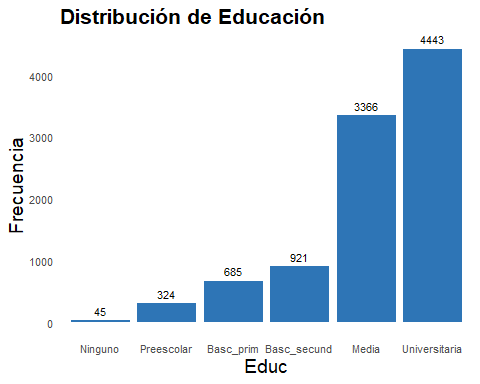


## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 27.00 34.00 36.44 45.00 86.00



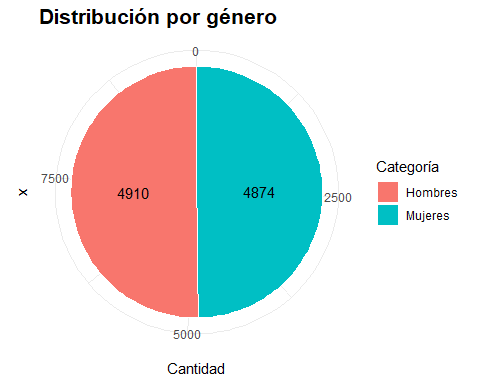
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 1 1 1 1 1

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 6.000 6.000 6.098 7.000 7.000

## educ  
## 1 3 4 5 6 7   
## 45 324 685 921 3366 4443



## sex  
## 0 1   
## 4875 4909

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 1.0000 0.5017 1.0000 1.0000

# # (7) Descripción de la variable de experiencia  
# expp<- DGEIH$exp  
# View(expp)  
# class(expp) #Aquí vemos la clase y nos dice que es numérica  
# mean(expp) #Aquí vemos la experiencia media y vemos es de 4.2 años  
# min(expp) #Aquí vemos la experiencia mínima y vemos que como era de esperar es de 0 años  
# max(expp) #Aquí vemos la experiencia máxima y vemos que es de 60 años  
# modeExp <- function(expp){  
# return(as.numeric(names(which.max(table(expp)))))  
# }  
# modeExp(expp) #Aquí vemos que la moda de la experiencia es de 0 años  
#   
#   
#   
# ###########  
#   
# # Crear el data frame  
# datos\_exp <- data.frame(expp)  
#   
# # Crear el histograma  
# histograma <- ggplot(datos\_exp, aes(x = expp)) +  
# geom\_histogram(binwidth = 1, fill = "#2E75B6", color = "white") +  
# labs(x = "Experiencia", y = "Frecuencia", title = "Histograma de Experiencia") +  
# theme\_minimal() +  
# theme(plot.title = element\_text(size = 16, face = "bold"),  
# axis.title = element\_text(size = 14),  
# axis.text = element\_text(size = 12))  
#   
# # Mostrar el histograma  
# print(histograma)  
#   
#   
#   
# ###########  
#   
#   
# plot(hist(expp))  
#   
#   
#   
# # (8) Descripción de la variable de oficio  
# library(modeest)  
# Oficio\_<- DGEIH$oficio  
# class(Oficio\_)  
# levels(Oficio\_)  
# summary(Oficio\_)  
# table(Oficio\_)  
# barplot(table((Oficio\_)))  
# mlv(Oficio\_, method = "mfv")  
#

## Age-wage profile

A great deal of evidence in Labor economics suggests that the typical worker’s age-wage profile has a predictable path: “Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50”

### Estimate the Age-wage profile

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

# Notas

Sustentar todas las medidas que salgan en las regresiones Exportar salidas de regresión con stargazer