Big Data, Machine Learning & Business Intelligence

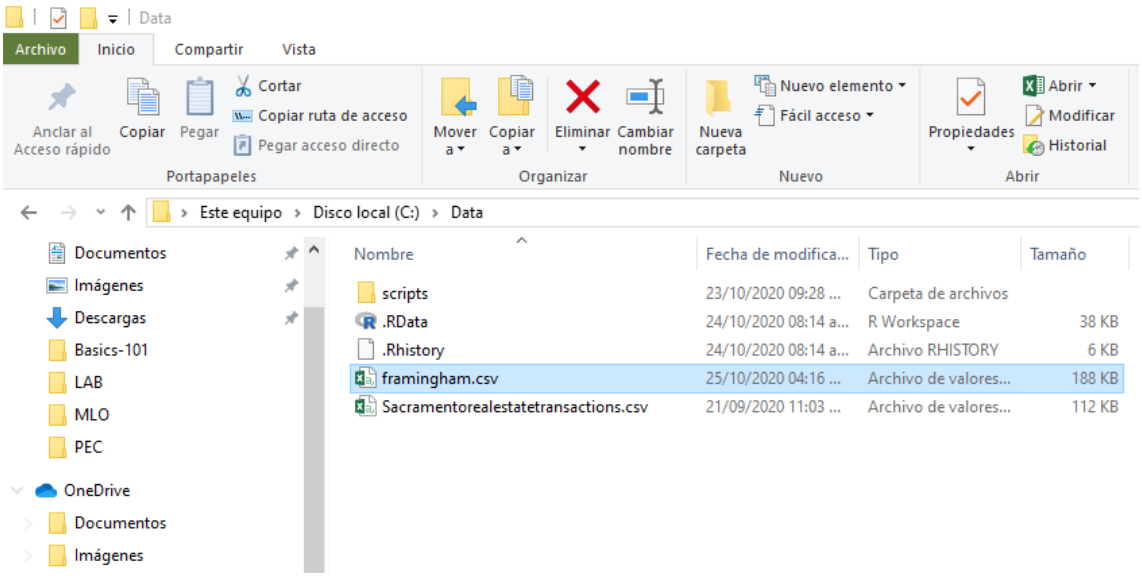
Lab: Implementar la Regresión Logística en Lenguaje R

Objetivos

- Implementar el algoritmo de machine learning supervisado regresión logística utilizando el lenguaje R

Procedimiento

1. Descargar el archivo framingham.csv, del repositorio GitHub del curso y copiarlo en C:/Data



2. Importa los datos de framingham.csv en un data frame, abre un archivo de script en R-Studio y ejecuta los siguientes comandos.

```
setwd("c:/Data")
ds = read.csv("framingham.csv")
str(ds)
```

3. Instala la librería caTools si es que esta aun no esta instalada.

```
install.packages("caTools")
```

4. Carga la librería caTools y establece el valor de la semilla del proceso aleatorio.

```
library(caTools)
set.seed(1000)
```

5. Divide el conjunto de datos en dos particiones una para entrenamiento y otra para pruebas, 70% y 30% respectivamente.

```
split = sample.split(datos$TenYearCHD, SplitRatio = 0.70)
train = subset(datos, split == TRUE)
test = subset(datos, split == FALSE)
particiones=c(nrow(train),nrow(test))
particiones
```

6. Crea el modelo de clasificación utilizando el algoritmo de regresión logística.

```
framinghamLog = glm(TenYearCHD ~ ., data = train, family = binomial(link = "logit"))
summary(framinghamLog)
```

7. Realiza a prueba de significancia del modelo.

   H_0: El modelo no es significativo
   H_1: El modelo es significativo.

   Se rechaza H_0 si $\propto > Valor\ P$

```
alfa = with(framinghamLog,null.deviance-deviance)
valor_P = with(framinghamLog,pchisq(alfa,df.null-df.residual,lower.tail = FALSE))
print(valor_P)
```

8. ¿El modelo es significativo?

   _____

9. Calcula la exponencial de los coeficientes. Interpreta cada uno. ¿Qué variable influye más en tener una enfermedad cardiovascular? ¿Qué variables influyen en forma positiva y negativa respecto al objetivo?

   _____
   _____
   _____