# Práctica de Evaluación FAD - Métodos de Análisis de Datos

Isabela Ignacio, Luisa Yánez, Miguel García

18/12/2021

## 0. Introducción

La práctica consiste en la elaboración y presentación de un informe de un proyecto de Ciencia de Datos, utilizando las técnicas aprendidas durante el curso, aplicadas a los datos seleccionados.

## 1. Uso de herramienta de control de versiones

El grupo eligió trabajar en lenguage R (RStudio version 1.4.1717) y utilizar como herramienta de control de versiones Github. El proyecto "/practica_fd_final" fue creado por Luisa Yánez (usuario lyanezgu) y compartido con los restantes participantes del grupo Isabela Ignacio (usuario IsaPires1329) y Miguel García (usuario mgarciasanc2021).

## 2. Análisis exploratório inicial - conjunto de datos

El conjunto de datos elegido por el grupo se llama "Hospital Charges in America" y incluye información que compara las tarifas de los servicios de hospitalizacón en diferentes estados de los EEUU para los 100 principales grupos de diagnósticos.

**Link del data set:** https://www.kaggle.com/dhirajnirne/hospital-charges-in-america.

### 2.1 Paquetes

```
library(readr)
library(ggplot2)
library(GGally)
library(dplyr)
library(tidyr)
library(missForest)
```

### 2.2 Cargar los datos

El conjunto de datos "Hospital Charges in America" contiene 12 columnas y 163065 filas y está en formato .csv. Inicialmente se guardó los datos en un data frame "hospital_charges" y hizo un estudio inicial de su contenido utilizando la función head y summary.

```
hospital_charges <- read_csv("notebooks/hospital-charges.csv")
head(hospital_charges)
```

```
## # A tibble: 6 x 12
##   `DRG Definition`         `Provider Id` `Provider Name`      `Provider Street ~
##   <chr>                            <dbl> <chr>                <chr>
## 1 039 - EXTRACRANIAL PROC~         10001 SOUTHEAST ALABAMA M~ 1108 ROSS CLARK C~
## 2 039 - EXTRACRANIAL PROC~         10005 MARSHALL MEDICAL CE~ 2505 U S HIGHWAY ~
## 3 039 - EXTRACRANIAL PROC~         10006 ELIZA COFFEE MEMORI~ 205 MARENGO STREET
## 4 039 - EXTRACRANIAL PROC~         10011 ST VINCENT'S EAST    50 MEDICAL PARK E~
## 5 039 - EXTRACRANIAL PROC~         10016 SHELBY BAPTIST MEDI~ 1000 FIRST STREET~
## 6 039 - EXTRACRANIAL PROC~         10023 BAPTIST MEDICAL CEN~ 2105 EAST SOUTH B~
## # ... with 8 more variables: Provider City <chr>, Provider State <chr>,
## #   Provider Zip Code <dbl>, Hospital Referral Region Description <chr>,
## #   Total Discharges <dbl>, Average Covered Charges <chr>,
## #   Average Total Payments <chr>, Average Medicare Payments <chr>
```

```
summary(hospital_charges)
```

```
##  DRG Definition      Provider Id     Provider Name      Provider Street Address
##  Length:163065      Min.   : 10001   Length:163065      Length:163065
##  Class :character   1st Qu.:110092   Class :character   Class :character
##  Mode  :character   Median :250007   Mode  :character   Mode  :character
##                     Mean   :255570
##                     3rd Qu.:380075
##                     Max.   :670077
##  Provider City      Provider State     Provider Zip Code
##  Length:163065      Length:163065      Min.   : 1040
##  Class :character   Class :character   1st Qu.:27261
##  Mode  :character   Mode  :character   Median :44309
##                                        Mean   :47938
##                                        3rd Qu.:72901
##                                        Max.   :99835
##  Hospital Referral Region Description Total Discharges  Average Covered Charges
##  Length:163065                        Min.   :  11.00   Length:163065
##  Class :character                     1st Qu.:  17.00   Class :character
##  Mode  :character                     Median :  27.00   Mode  :character
##                                       Mean   :  42.78
##                                       3rd Qu.:  49.00
##                                       Max.   :3383.00
##  Average Total Payments Average Medicare Payments
##  Length:163065          Length:163065
##  Class :character       Class :character
##  Mode  :character       Mode  :character
##
##
##
```

## 2.2 Definición de las variables que componen los datos de estudio

Las 12 variables que componen los datos pueden ser descritas como:

- **DRG Definition:** Grupo relativo a un diagnostico. Los grupos de diagnóstico relacionado (DRG) se utilizan para clasificar la gravedad de la enfermedad en las visitas hospitalarias de pacientes hospitalizados, el riesgo de mortalidad, el pronóstico, la dificultad del tratamiento, la necesidad de intervención y la intensidad de los recursos que necesitan. El sistema DRG fue desarrollado en la Universidad de Yale en la década de 1970 para la clasificación estadística de casos hospitalarios. Realmente la variable DRG es relativa al código y la descripción que identifican el MS-DRG. Los MS-DRG son un sistema de clasificación que agrupa condiciones clínicas similares (diagnósticos) y los procedimientos proporcionados por el hospital durante la estancia. Luego estamos hablando de categorías de estadías hospitalarias para pacientes hospitalizados. El sistema de Medicare (Sistema de Seguridad Social en EEUU) los utiliza para determinar los reembolsos para hospitales, centros de enfermería especializada y hospicios. Una estadía en el hospital puede variar de un día a 100 días. Los MS-DRG más caros tienen las estadías promedio más largas. El establecimiento del cada DRG se establece según las condiciones clínicas del paciente, necesidad de cantidades similares de recursos para pacientes hospitalizados y sexo y edad del paciente. Para ello se utiliza el sistema de DRG llamado "Medicare Severity DRGs (MS-DRGs)" para reflejar en mejor manera la severidad de la enfermedad del paciente y su consumo de recursos para su recuperación. Para clasificar el nivel de severidad de un paciente dentro del sistema "MS-DRGs" hay códigos secundarios de diagnóstico:
  - MCC: Major Complication/Comorbidity -> El nivel más alto de severidad.
  - CC: Complication/Comorbidity -> El siguiente nivel de severidad.
  - Non-CC: Non-Complication/Comorbidity -> Este nivel no supone una gran severidad en la enfermedad ni un gran gasto de recursos;
- **Provider ID:** ID o número identificativo de referencia del hospital;
- **Provider Name:** Nombre del hospital;
- **Provider Street Address:** Dirección postal donde se ubica el hospital;
- **Provider City:** Ciudad donde se ubica el hospital;
- **Provider State:** Estado federal de EEUU donde se ubica el hospital;
- **Provider Zip Code:** Código postal donde se ubica el hospital;
- **Hospital Referral Region Description:** Delinación geográfica específica creada por la organización norteamericana "Dartmouth Atlas of Health Care", para estudiar los mercados vinculados al sector salud en EEUU;
- **Total Discharges:** Número de personas dadas de alta;
- **Average Covered Charges:** Gastos medios del hospital por los servicios cubiertos por la seguridad social para todas las altas del grupo relacionado con el diagnóstico. Por lo tanto cargo promedio según grupo de diagnóstico DRG establecido. Los pacientes que tienen características clínicas similares y costos de tratamiento similares se asignan a un Grupo de Diagnóstico Relacionado (DRG). El DRG está vinculado a un monto de pago fijo basado en el costo promedio del tratamiento de los pacientes del grupo.La asignación de DRG se basa en el diagnóstico del paciente, los procedimientos recibidos, la edad y otra información. Por lo tanto esta variable contiene el cargo promedio por cada DRG proporcionado por el hospital. Sus cargos promedio podrían ser más o menos dependiendo de las necesidades específicas de su paciente y los servicios prestados.E sto es lo que el hospital cobra en la factura final del hospital y es equivalente al "sticker price". Este es en gran medida un número irrelevante, ya que no importa lo que cobren los diferentes hospitales, a todos se les pagará la misma cantidad de Medicare por cualquier DRG dado. Prácticamente nadie paga el "stiker price" en un hospital. Cuando un paciente ha sido admitido como hospitalizado en un hospital, ese hospital asigna un DRG cuando este paciente es dado de alta, basándolo en la atención que necesitaba durante su estadía en el hospital. Al hospital se le paga una cantidad fija por ese DRG, independientemente de cuánto dinero realmente gaste en su tratamiento. Si un hospital puede tratar a un paciente de forma

efectiva por menos dinero del que Medicare paga por su DRG, entonces el hospital gana dinero con esa hospitalización. Si el hospital gasta más dinero cuidando del paciente de lo que Medicare le da para su DRG, entonces el hospital pierde dinero en esa hospitalización;

- **Average Medicare Payments:**Importe medio cubierto por la Seguridad Social de EEUU. Esto es lo que Medicare paga al hospital por ese DRG;
- **Average Total Payments:** Importe medio total a pagar por persona. Esto es lo que realmente se le paga al hospital e incluye lo que paga Medicare más los copagos que paga el paciente más cualquier cosa que pague el seguro secundario (seguro privado).

# 3. Definición de objetivos

El objetivo final del proyecto es llegar a un modelo para recomendar que hospital debe elegir un paciente enfermo en EEUU, en base a su posible enfermedad, su localización y los costes que su caso clínico puede legar a tener.

Para esta primera entrega el objetivo es realizar el tratamiento de datos adecuado y seleccionar las mejores variables que servirán para llegar al modelo Machine Learning desedo.Se desea también hacer un ajuste, interpretación y diagnosis del modelo de regresión linela múltiple.

# 4. Transformaciones de variables cuantitativas y procesado de variables cualitativas - Limpieza de datos

Se ha decidido cambiar los nombres de las columnas para seguir un patrón y eliminar el simbolo de dolar de las últimas tres columnas, transformando las columnas a tipo numérico.

## 4.1 Cambiar los nombres de las columnas

```
names(hospital_charges) <- c("drg_def", "prov_id", "prov_name",
    "prov_address", "prov_city", "prov_state", "prov_zip", "referral_reg",
    "total_discharges", "mean_covered_charges", "mean_total_payments",
    "mean_medicare_payments")
head(hospital_charges)
```

```
## # A tibble: 6 x 12
##   drg_def      prov_id prov_name     prov_address  prov_city prov_state prov_zip
##   <chr>          <dbl> <chr>         <chr>         <chr>     <chr>         <dbl>
## 1 039 - EXTRA~   10001 SOUTHEAST AL~ 1108 ROSS CL~ DOTHAN    AL            36301
## 2 039 - EXTRA~   10005 MARSHALL MED~ 2505 U S HIG~ BOAZ      AL            35957
## 3 039 - EXTRA~   10006 ELIZA COFFEE~ 205 MARENGO ~ FLORENCE  AL            35631
## 4 039 - EXTRA~   10011 ST VINCENT'S~ 50 MEDICAL P~ BIRMINGH~ AL            35235
## 5 039 - EXTRA~   10016 SHELBY BAPTI~ 1000 FIRST S~ ALABASTER AL            35007
## 6 039 - EXTRA~   10023 BAPTIST MEDI~ 2105 EAST SO~ MONTGOME~ AL            36116
## # ... with 5 more variables: referral_reg <chr>, total_discharges <dbl>,
## #   mean_covered_charges <chr>, mean_total_payments <chr>,
## #   mean_medicare_payments <chr>
```

### 4.2 Eliminar dolar de las ultimas 3 columnas

```
hospital_charges$mean_covered_charges = as.numeric(gsub("\\$",
    "", hospital_charges$mean_covered_charges))

hospital_charges$mean_total_payments = as.numeric(gsub("\\$",
    "", hospital_charges$mean_total_payments))

hospital_charges$mean_medicare_payments = as.numeric(gsub("\\$",
    "", hospital_charges$mean_medicare_payments))

head(hospital_charges)
```

```
## # A tibble: 6 x 12
##   drg_def      prov_id prov_name     prov_address  prov_city prov_state prov_zip
##   <chr>          <dbl> <chr>         <chr>         <chr>     <chr>         <dbl>
## 1 039 - EXTRA~   10001 SOUTHEAST AL~ 1108 ROSS CL~ DOTHAN    AL            36301
## 2 039 - EXTRA~   10005 MARSHALL MED~ 2505 U S HIG~ BOAZ      AL            35957
## 3 039 - EXTRA~   10006 ELIZA COFFEE~ 205 MARENGO ~ FLORENCE  AL            35631
## 4 039 - EXTRA~   10011 ST VINCENT'S~ 50 MEDICAL P~ BIRMINGH~ AL            35235
## 5 039 - EXTRA~   10016 SHELBY BAPTI~ 1000 FIRST S~ ALABASTER AL            35007
## 6 039 - EXTRA~   10023 BAPTIST MEDI~ 2105 EAST SO~ MONTGOME~ AL            36116
## # ... with 5 more variables: referral_reg <chr>, total_discharges <dbl>,
## #   mean_covered_charges <dbl>, mean_total_payments <dbl>,
## #   mean_medicare_payments <dbl>
```

### 4.3 Creando columna nueva mean_tota_payments - mean_medicare_payments

Nueva variable con el valor de los copagos que paga el paciente más cualquier cosa que pague el seguro secundario (seguro privado).

## 5. Detección, tratamiento e imputación de datos faltantes

Através de la función summary se comprobó que no hay datos faltantes en el data set y por eso el grupo tuvo que añadirlos manualmente para aproximarse mejor de un caso real con datos faltantes. Los datos faltantes fueron imputados solo en las columnas que no van a servir de analísis para este estudio.

Uctilizamos la libreria missForest y generamos una semilla para que el resultado sea siempre el mismo.

```
set.seed(1)
hospital_charges <- bind_cols(hospital_charges[c(1, 3, 5, 7,
    8, 9, 11, 12)], missForest::prodNA(hospital_charges[c(-1,
    -3, -5, -7, -8, -9, -11, -12)], noNA = 0.1))

hospital_charges
```

```
## # A tibble: 163,065 x 12
##    drg_def       prov_name      prov_city prov_zip referral_reg total_discharges
##    <chr>         <chr>          <chr>        <dbl> <chr>                   <dbl>
##  1 039 - EXTRAC~ SOUTHEAST ALA~ DOTHAN       36301 AL - Dothan                91
```

```
##  2 039 - EXTRAC~ MARSHALL MEDI~ BOAZ         35957 AL - Birmin~               14
##  3 039 - EXTRAC~ ELIZA COFFEE ~ FLORENCE     35631 AL - Birmin~               24
##  4 039 - EXTRAC~ ST VINCENT'S ~ BIRMINGH~    35235 AL - Birmin~               25
##  5 039 - EXTRAC~ SHELBY BAPTIS~ ALABASTER    35007 AL - Birmin~               18
##  6 039 - EXTRAC~ BAPTIST MEDIC~ MONTGOME~    36116 AL - Montgo~               67
##  7 039 - EXTRAC~ EAST ALABAMA ~ OPELIKA      36801 AL - Birmin~               51
##  8 039 - EXTRAC~ UNIVERSITY OF~ BIRMINGH~    35233 AL - Birmin~               32
##  9 039 - EXTRAC~ HUNTSVILLE HO~ HUNTSVIL~    35801 AL - Huntsv~              135
## 10 039 - EXTRAC~ GADSDEN REGIO~ GADSDEN      35903 AL - Birmin~               34
## # ... with 163,055 more rows, and 6 more variables: mean_total_payments <dbl>,
## #   mean_medicare_payments <dbl>, prov_id <dbl>, prov_address <chr>,
## #   prov_state <chr>, mean_covered_charges <dbl>
```

# 6. Partición del conjunto de datos: data set training y data set test

Se divide el conjunto de datos en dos (20% test y 80% training), guardando la partición test para ser utilizada par la validación del modelo final y trabajando con la partición training.

```
set.seed(101)
sample <- sample.int(n = nrow(hospital_charges), size = floor(0.8 *
    nrow(hospital_charges)), replace = F)
train <- hospital_charges[sample, ]
test <- hospital_charges[-sample, ]

train
```

```
## # A tibble: 130,452 x 12
##    drg_def         prov_name    prov_city prov_zip referral_reg total_discharges
##    <chr>           <chr>        <chr>        <dbl> <chr>                    <dbl>
##  1 064 - INTRACRA~ RIVERVIEW H~ NOBLESVI~    46060 IN - Indian~                11
##  2 439 - DISORDER~ ST LUKE'S R~ NEW YORK     10025 NY - Manhat~                13
##  3 853 - INFECTIO~ ST JOSEPH'S~ YONKERS      10701 NY - White ~                17
##  4 329 - MAJOR SM~ UNIVERSITY ~ KANSAS C~    66160 MO - Kansas~                44
##  5 195 - SIMPLE P~ GARDEN CITY~ GARDEN C~    48135 MI - Dearbo~                52
##  6 176 - PULMONAR~ HORIZON MED~ DICKSON      37055 TN - Nashvi~                14
##  7 641 - MISC DIS~ BAYLOR UNIV~ DALLAS       75246 TX - Dallas                110
##  8 638 - DIABETES~ ST ELIZABET~ FLORENCE     41042 KY - Coving~                15
##  9 872 - SEPTICEM~ ST JOSEPH'S~ SYRACUSE     13203 NY - Syracu~                70
## 10 439 - DISORDER~ SOUTH POINT~ WARRENSV~    44122 OH - Clevel~                20
## # ... with 130,442 more rows, and 6 more variables: mean_total_payments <dbl>,
## #   mean_medicare_payments <dbl>, prov_id <dbl>, prov_address <chr>,
## #   prov_state <chr>, mean_covered_charges <dbl>
```

```
test
```

```
## # A tibble: 32,613 x 12
##    drg_def       prov_name      prov_city prov_zip referral_reg total_discharges
##    <chr>         <chr>          <chr>        <dbl> <chr>                    <dbl>
##  1 039 - EXTRAC~ MARSHALL MEDI~ BOAZ         35957 AL - Birmin~                14
##  2 039 - EXTRAC~ SOUTH BALDWIN~ FOLEY        36535 AL - Mobile                 15
##  3 039 - EXTRAC~ MOBILE INFIRM~ MOBILE       36652 AL - Mobile                 66
##  4 039 - EXTRAC~ TUCSON MEDICA~ TUCSON       85712 AZ - Tucson                 40
```

```
##  5 039 - EXTRAC~ CARONDELET ST~ TUCSON       85711 AZ - Tucson                42
##  6 039 - EXTRAC~ ST JOSEPH'S H~ PHOENIX      85013 AZ - Phoenix               18
##  7 039 - EXTRAC~ BANNER BOSWEL~ SUN CITY     85351 AZ - Sun Ci~               62
##  8 039 - EXTRAC~ SUMMIT HEALTH~ SHOW LOW     85901 AZ - Phoenix               17
##  9 039 - EXTRAC~ BANNER HEART ~ MESA         85206 AZ - Mesa                  64
## 10 039 - EXTRAC~ CONWAY REGION~ CONWAY       72034 AR - Little~               17
## # ... with 32,603 more rows, and 6 more variables: mean_total_payments <dbl>,
## #   mean_medicare_payments <dbl>, prov_id <dbl>, prov_address <chr>,
## #   prov_state <chr>, mean_covered_charges <dbl>
```