INSTITUTO TECNOLÓGICO DE BUENOS AIRES - ITBA
ESCUELA DE INNOVACIÓN

TRABAJO PRÁCTICO 1

AUTORES: Lambertucci, Guido Enrique (Leg. Nº 58009) DNI: 40397224

DOCENTES: Riccillo, Marcela Leticia

IA.03 - Inteligencia Artificial

BUENOS AIRES

Índice

1. Ejercicio 1

1.1. Consigna

- Primera Parte Ejercicio 1 – Regresión Los casos de Regresión se caracterizan por tener una variable cuantitativa para predecir.
  - Seleccione un dataset con un caso de Regresión. El dataset debe ser obtenido de alguna librería de R o de una página web pública (no incluir datos confidenciales).
  - Por ejemplo, se podría utilizar:
    $\Longrightarrow$ Datasets de R como: mtcars de base, iris de base, cheddar de faraway, etc.
    $\Longrightarrow$ datasets de UCI (Universidad de California) `https://archive.ics.uci.edu`
    $\Longrightarrow$ datasets de Kaggle `https://www.kaggle.com/`
    $\Longrightarrow$ datasets de ISLR `https://www.statlearning.com/resources-second-edition`
  - El dataset debe contener al menos 3 variables y una de ellas debe ser numérica. (Nota: este dataset es solamente para este ejercicio y no se espera ser utilizado en otros ejercicios).
    1. Indique el nombre del dataset, y la librería de R o la página web fuente del mismo.
    2. ¿De qué trata la base?
    3. ¿Cuántos registros tiene la base? ¿Cuántas variables? ¿De qué tipo son las variables? Podría utilizar `dim(base)`, `str(base)`, `summary(base)`
    4. Realice un histograma de la variable numérica seleccionada. ¿En qué rango se encuentran los valores? `hist(variable, main="Título", col="color")`
       a) Para el título ingrese su nombre, como “Histograma de Marcela”.
       b) Elija un color para el gráfico. Tenga en cuenta que ingresando `colors()` en R verá que hay más de 500 colores posibles.
       c) Indique el código R utilizado.

1.2. Resolución

La base de datos elegida es Summer Olympics Weightlifting records 2000 to 2020 de la página web Kaggle. Esta base registra los pesos máximos levantados por cada atleta olímpico en las dos movimientos correspondientes al levantamiento de pesas (Clean & Jerk y Snatch), ademas se provee el genero del atleta, su peso corporal, entre otras cosas. Cuenta con 716 registros, con 11 variables (1).

Nombre de la variable	Tipo	Comentario
x	Integer	Indice
Athlete	String	Nombre del Atleta
Bodyweight..kg	Double	Peso corporal en kilogramos
Clean...Jerk..kg	Double	Peso del Clean & Jerk
Snatch..kg	Double	Peso del Snatch
Total..kg	Double	Peso total
Ranking	Integer	Clasificaci\'on final del atleta
Url	String	Link a wikipedia
Title	String	Titulo obtenido
Year	Integer	Año de competición
Gender	String	Género

Tabla 1: Tabla de variables.

De estas variables, dependiendo que es lo que se quiere predecir, hay varias que no aportan información alguna. Por ejemplo si uno quisiera predecir el valor máximo de Snatch la variable x no aporta informacion, al igual que la Url. En contraposición el Clean & Jerk y el peso corporal son variables significativas para esta predicción. Se optó por la variable "Snatch" para graficar (1).

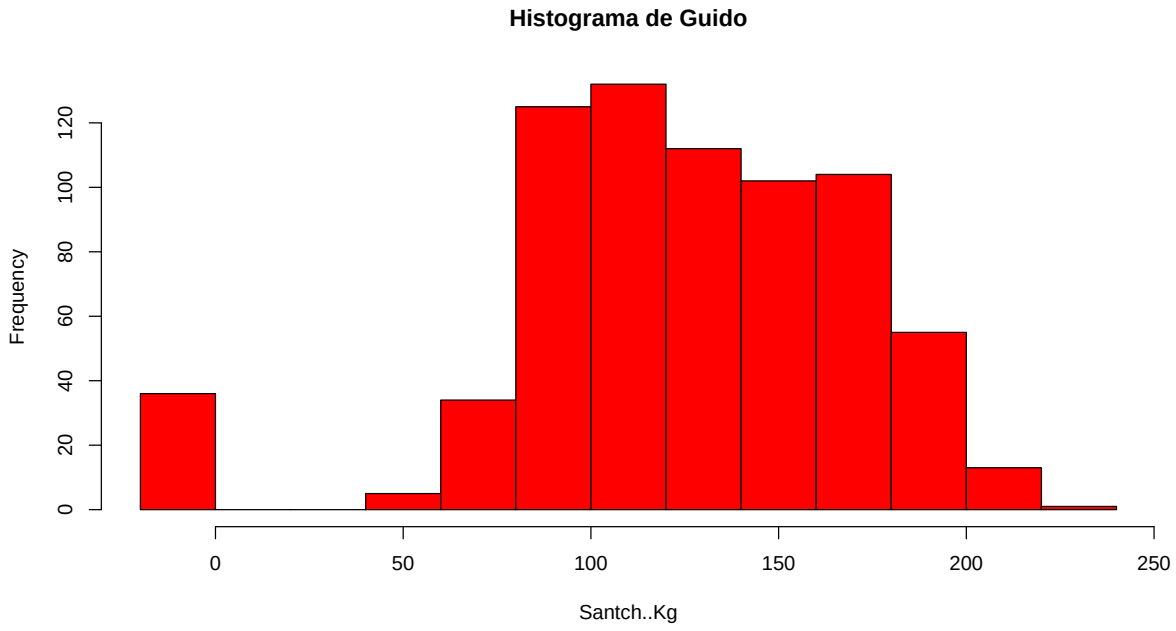

Figura 1: Histograma de la variable "Snatch..Kg".

Es interesante observar el histograma, ya que revela dos aspectos distintos. Por un lado, se puede apreciar una distribución centrada alrededor de los 125 Kg, la cual muestra una forma de campana similar a una distribución normal. Por otro lado, se destacan una serie de valores en 0 Kg. Estos valores atípicos, conocidos como "outliers", no son representativos del rendimiento del atleta ni contribuyen a su predicción. Más bien, reflejan casos de atletas que no compitieron en esa categoría y se les asignó un valor de 0.

1.3. Código utilizado

```
#install.packages("rstudioapi") #Install rstudioapi for the automatic set
    working directory feature
library(caret)

directory <- dirname(rstudioapi::getActiveDocumentContext()$path)
#just use the setwd if not using rstudio
setwd(directory)
base=read.table("./base.csv",sep=",",header=TRUE)#Loading dataset
#Some usefull information for getting to know my dataset
dim(base)
str(base)
summary(base)
View(base)

hist(base$Snatch..kg.,main = "Histograma de Guido",xlab = "Santch..Kg",
     ylab = "Frequency",col = "red")
```