Actividad 001: Fundamentos de Python y RMarkdown

Ph. D. Pablo Eduardo Caicedo R.

6 de octubre de 2021

Contenidos

1 Conjunto de datos

En el año 1936, Ronald Fisher publica su artículo titulado "The use of multiple measurements in taxonomic problems" donde ejemplifica la técnica estadística *análisis lineal discriminante*. Para ello utiliza un conjunto de datos; colectado por Edgar Anderson, el cual tiene información de mediciones de 150 flores de la familia iridacea, 50 de la especie Iris setosa, 50 de la especie Iris virginica y 50 de la especie Iris versicolor.

El conjunto de datos posee 5 características: ancho y largo de sépalo, al igual que ancho y largo del pétalo. Finalmente tiene una columna de clase donde se encuentra la especie de la flor.

1.1 Ubicación del archivo

El dataset puede ser fácilmente encontrado en internet, por ejemplo en el repositorio del Centro para el Aprendizaje de Máquina y Sistemas Inteligentes de la Universidad de California; el cual se encuentra en esta url.

La descarga y la descripción completa del DATASET se encuentra en la siguiente url.

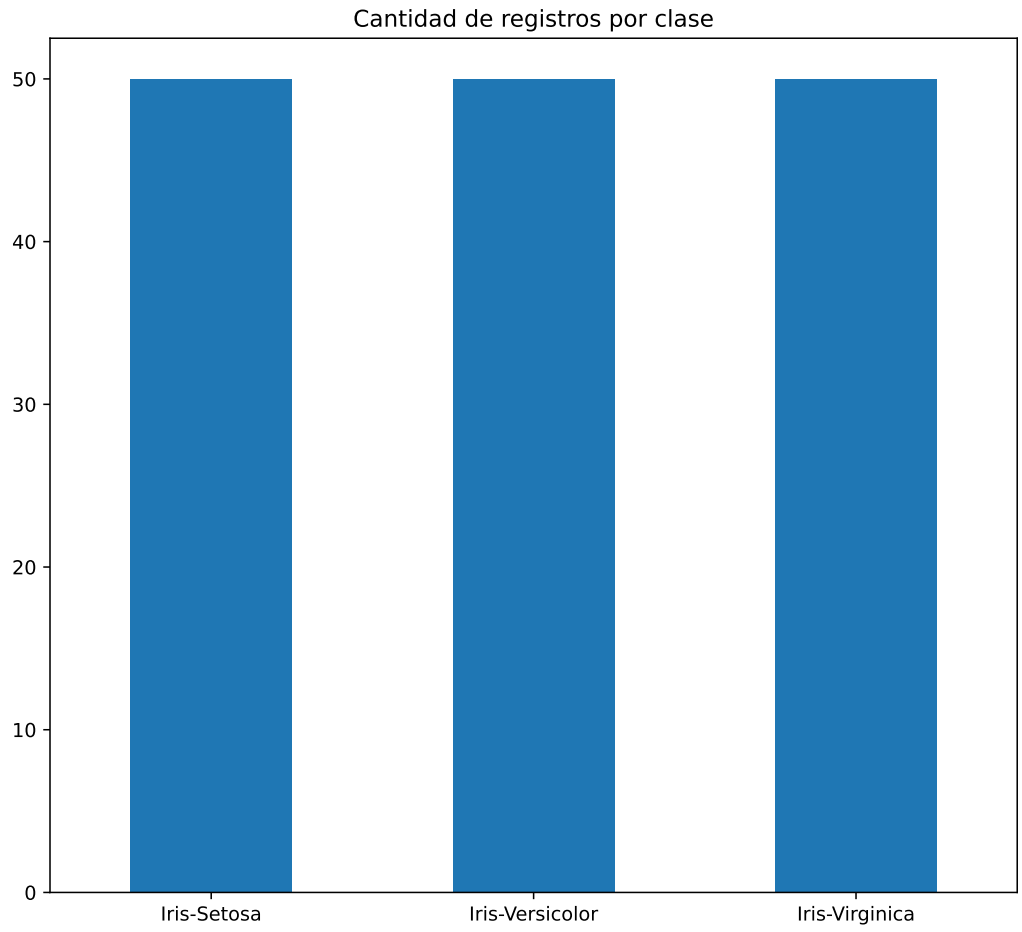
1.2 El conjunto de datos en Python

Sin embargo, la forma más sencilla de utilizarlo es através de la librería *Scikit-learn*, que es instalable vía conda con el comando:

Una vez se ha instalado la librería ya será utilizable en jupyter utilizando el siguiente código:

Utilizando el módulo *datasets* y la función *load.iris()* se cargan los datos del conjunto de datos

Finalmente, se hace una adecuación del formato del dataset y se realiza una gráfica básica de la información de la clase.



Se advierte que en la columna Target:

- 0 equivale a la especie Iris-Setosa,
- 1 a la especie Iris_Versicolor
- 2 a la especie Iris-Virginica

1.3 Estadística descriptiva del conjunto de datos

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

1.3.1 Tendencia central

1.3.1.1 Tendencia central para la especie Iris-Setosa

1.3.1.2 Tendencia central para la especie Iris-Versicolor

1.3.1.3 Tendencia central para la especie Iris-Virginica

1.3.2 Dispersión

1.3.2.1 Dispersión para la especie Iris-Setosa

1.3.2.2 Dispersión para la especie Iris-Versicolor

1.3.2.3 Dispersión para la especie Iris-Virginica