

Intro a Python 2020

Martín Palazzo

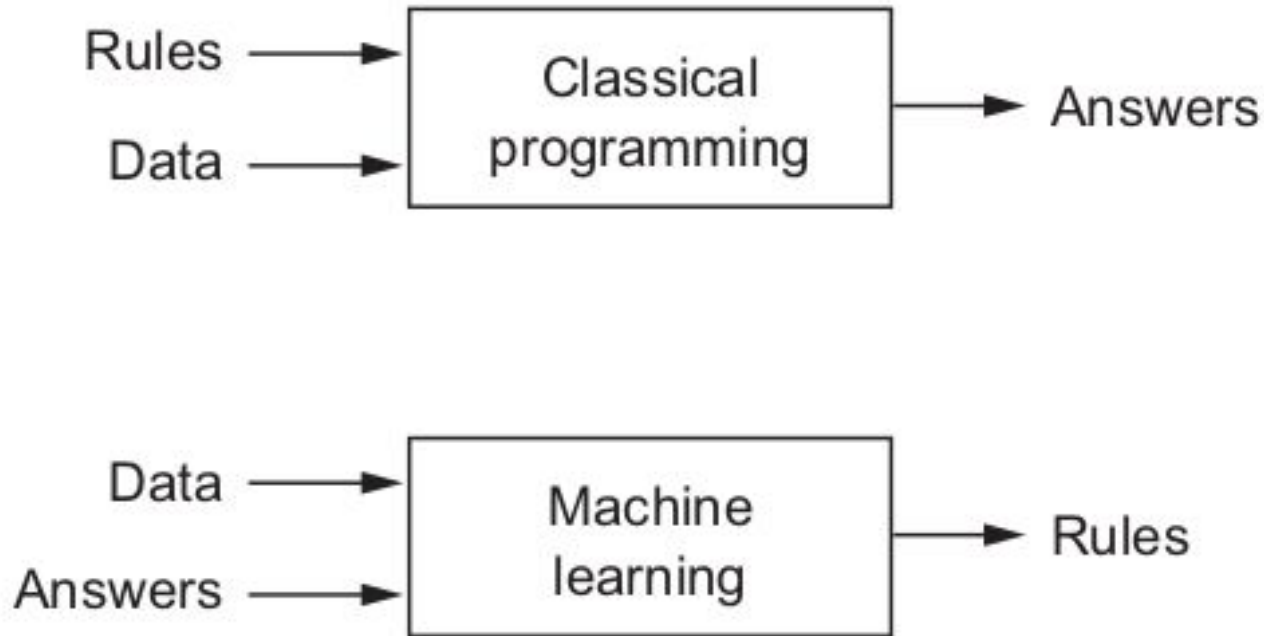
Biomedicine Research Institute of Buenos Aires

Université de Technologie de Troyes

Universidad Tecnológica Nacional BA



machine learning in a nutshell



Samples (instancia) & Features (atributos)

Los datos estarán caracterizados por dos indicadores:

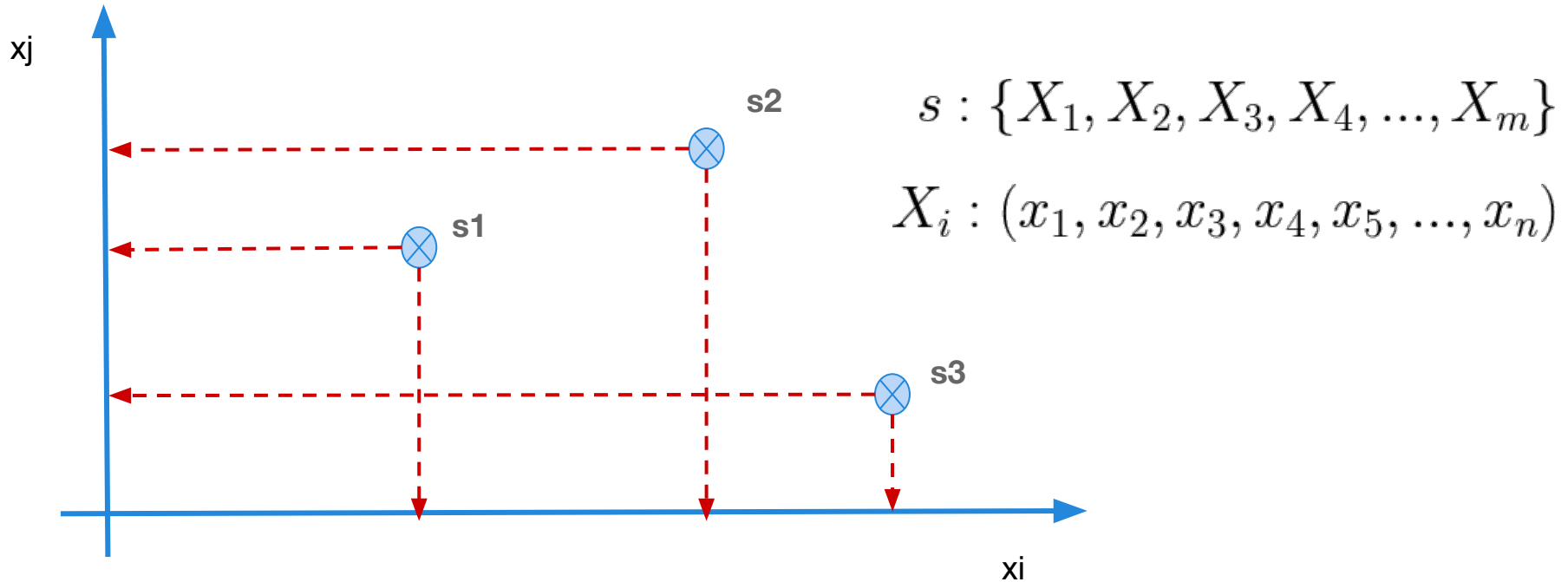
Samples

Las samples corresponden a las instancias que obtenemos de una **muestra** de datos. Dicha muestra pertenece a una población que generalmente no conocemos por completo. Nuestro set de datos tendrá una cantidad determinada de samples.

Features

Denominamos features (atributos o mediciones) a las **variables** que definen a cada sample (instancia). La cantidad de features que posea un sample es equivalente a la cantidad de **dimensiones** que describen a esa instancia en un espacio de alta dimensión. Nuestros datos “viven” en un espacio n-dimensional.

Samples (instancia) & Features (atributos)



¿Cuántas features y cuantas samples hay en este ejemplo?

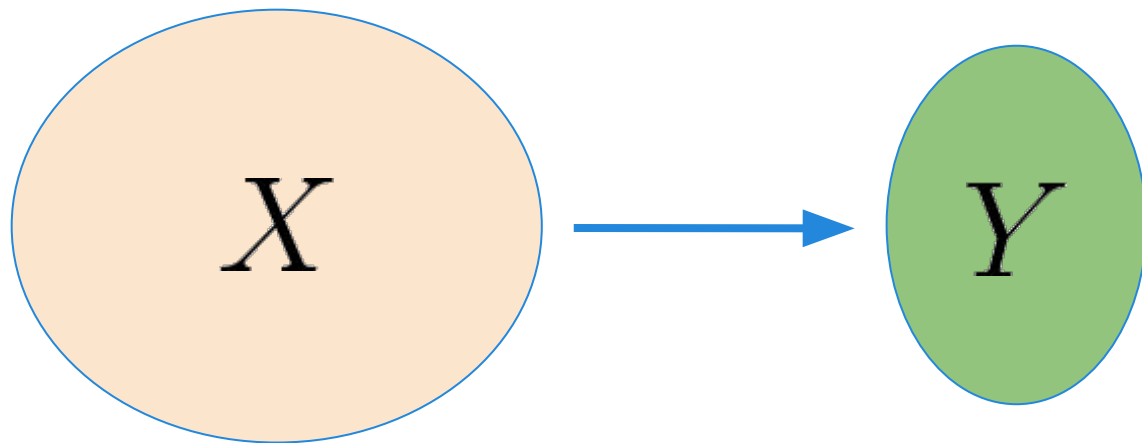
Aprendizaje Supervisado

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$$

Suponemos un dataset con observaciones/samples S , donde X_i es un vector de features e Y_i es una label (etiqueta) asociada a cada observación X_i .

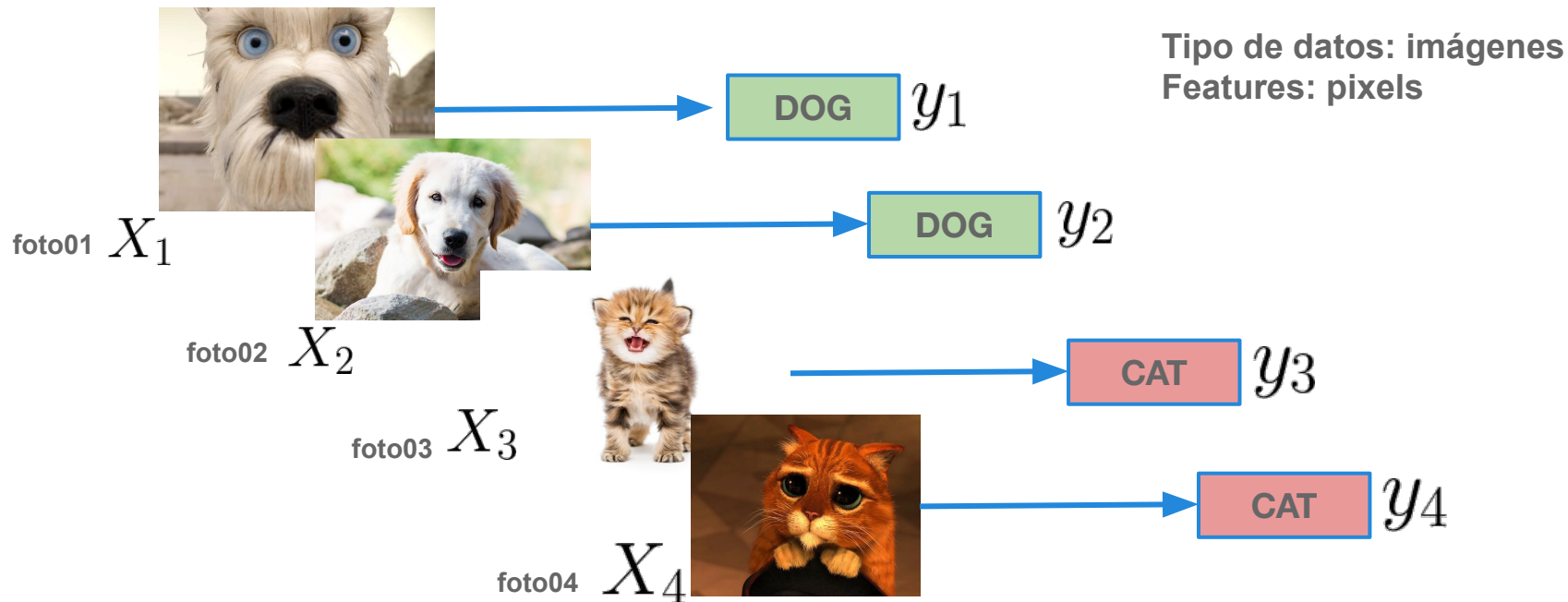
Solemos denominar a cada observación X_i como “sample” y a cada etiqueta Y_i como “label”.

Aprendizaje Supervisado



Suponemos que la variable Y es dependiente de X . Esto quiere decir que Y está condicionada y es consecuencia de X . **Lo que no conocemos es la función $y = f(x)$** y es $f(x)$ lo que queremos aprender desde los datos.

Tipos de Aprendizaje: Supervisado



Cada instancia (sample) viene acompañada de una etiqueta (label).

Tipos de Aprendizaje: Supervisado

Tipo de datos: ADN
Features: mutaciones

CONTROL SANO y_1

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=sequencingProgramV3.1
##reference=file:///seq/references/1000GenomePilot-NCBI36.fasta
##contig=ID=0,length=5436364,assembly=B36,md5=f126c02f86dc73776618f66b6b26,species="Homo sapiens",taxonomy=
##phasing=partial
##INFO=ID=0,Number=1,Type=Integer,Description="Number of Samples With Data"
##INFO=ID=1,Number=1,Type=Integer,Description="Total Depth"
##INFO=ID=4,Number=1,Type=Float,Description="Allele Frequency"
##INFO=ID=4,Number=1,Type=String,Description="Ancestral Allele"
##INFO=ID=8,Number=0,Type=Flag,Description="GDP membership, build 129"
##INFO=ID=10,Number=0,Type=Flag,Description="HighP membership"
##FILTER=ID=0,Description="Quality below 10"
##FILTER=ID=10,Description="Less than 50% of samples have data"
##FORMAT=ID=0,Number=1,Type=String,Description="Genotype"
##FORMAT=ID=0,Number=1,Type=String,Description="Genotype Quality"
##FORMAT=ID=0,Number=1,Type=String,Description="Read Depth"
##FORMAT=ID=0,Number=2,Type=Integer,Description="PairType Quality"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SRA0001 SRA0002 SRA0003
20 14370 rs604207 C A 20 PASS RS=0;DP=1;AF=0.0;DB;2 GT:DP:HQ 0:0:48:1:51,51 1:0:48:0:51,51 1/1:48:1...
20 17330 T A 3 GQ RS=0;DP=1;AF=0.0;GT:DP:HQ 0:0:49:3:58,50 0:1:3:1:65,3 0/0:41:3
20 110906 rs604035 A G,T 67 PASS RS=0;DP=13;AF=0.333,0.667;AA=T;DB GT:DP:HQ 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 130237 T 47 PASS RS=0;DP=13;AA=T GT:DP:HQ 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 1234567 microsat1 CTC G,CTCT 50 PASS RS=0;DP=9;AA=G GT:DP: 0/1:38:4 0/2:17:2 1/1:40:3
```

paciente01 X_1

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=sequencingProgramV3.1
##reference=file:///seq/references/1000GenomePilot-NCBI36.fasta
##contig=ID=0,length=5436364,assembly=B36,md5=f126c02f86dc73776618f66b6b26,species="Homo sapiens",taxonomy=
##phasing=partial
##INFO=ID=0,Number=1,Type=Integer,Description="Number of Samples With Data"
##INFO=ID=1,Number=1,Type=Integer,Description="Total Depth"
##INFO=ID=4,Number=1,Type=Float,Description="Allele Frequency"
##INFO=ID=4,Number=1,Type=String,Description="Ancestral Allele"
##INFO=ID=8,Number=0,Type=Flag,Description="GDP membership, build 129"
##INFO=ID=10,Number=0,Type=Flag,Description="HighP membership"
##FILTER=ID=0,Description="Quality below 10"
##FILTER=ID=10,Description="Less than 50% of samples have data"
##FORMAT=ID=0,Number=1,Type=String,Description="Genotype"
##FORMAT=ID=0,Number=1,Type=String,Description="Genotype Quality"
##FORMAT=ID=0,Number=1,Type=String,Description="Read Depth"
##FORMAT=ID=0,Number=2,Type=Integer,Description="PairType Quality"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SRA0001 SRA0002 SRA0003
20 14370 rs604207 C A 20 PASS RS=0;DP=1;AF=0.0;DB;2 GT:DP:HQ 0:0:48:1:51,51 1:0:48:0:51,51 1/1:48:1...
20 17330 T A 3 GQ RS=0;DP=1;AF=0.0;GT:DP:HQ 0:0:49:3:58,50 0:1:3:1:65,3 0/0:41:3
20 110906 rs604035 A G,T 67 PASS RS=0;DP=13;AF=0.333,0.667;AA=T;DB GT:DP:HQ 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 130237 T 47 PASS RS=0;DP=13;AA=T GT:DP:HQ 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 1234567 microsat1 CTC G,CTCT 50 PASS RS=0;DP=9;AA=G GT:DP: 0/1:38:4 0/2:17:2 1/1:40:3
```

CANCER

y_2

paciente02 X_2

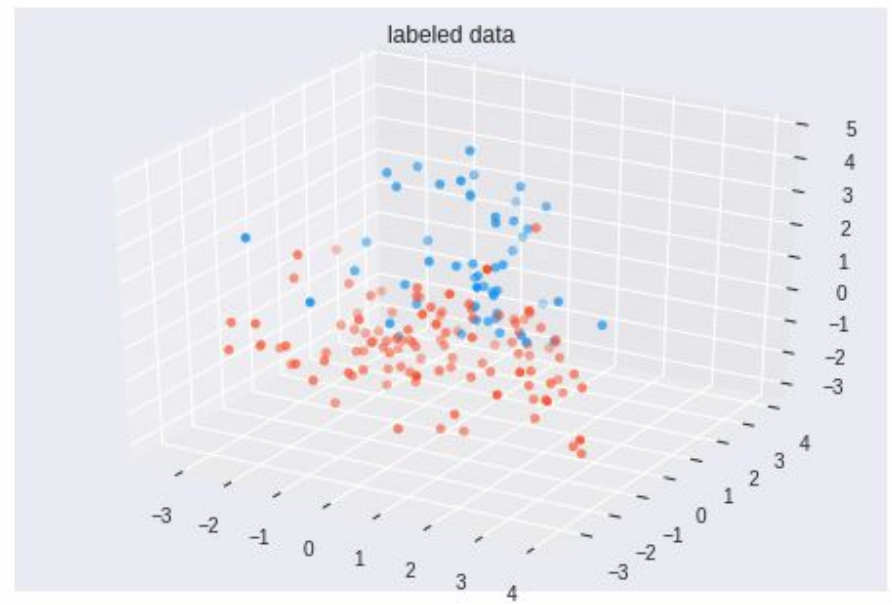
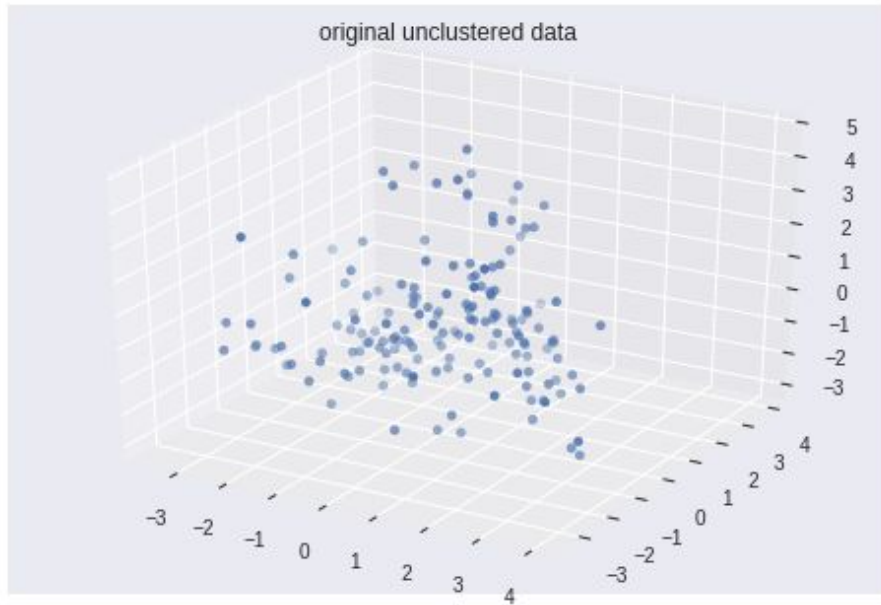
```
##fileformat=VCFv4.1
##fileDate=20090805
##source=sequencingProgramV3.1
##reference=file:///seq/references/1000GenomePilot-NCBI36.fasta
##contig=ID=0,length=5436364,assembly=B36,md5=f126c02f86dc73776618f66b6b26,species="Homo sapiens",taxonomy=
##phasing=partial
##INFO=ID=0,Number=1,Type=Integer,Description="Number of Samples With Data"
##INFO=ID=1,Number=1,Type=Integer,Description="Total Depth"
##INFO=ID=4,Number=1,Type=Float,Description="Allele Frequency"
##INFO=ID=4,Number=1,Type=String,Description="Ancestral Allele"
##INFO=ID=8,Number=0,Type=Flag,Description="GDP membership, build 129"
##INFO=ID=10,Number=0,Type=Flag,Description="HighP membership"
##FILTER=ID=0,Description="Quality below 10"
##FILTER=ID=10,Description="Less than 50% of samples have data"
##FORMAT=ID=0,Number=1,Type=String,Description="Genotype"
##FORMAT=ID=0,Number=1,Type=String,Description="Genotype Quality"
##FORMAT=ID=0,Number=1,Type=String,Description="Read Depth"
##FORMAT=ID=0,Number=2,Type=Integer,Description="PairType Quality"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SRA0001 SRA0002 SRA0003
20 14370 rs604207 C A 20 PASS RS=0;DP=1;AF=0.0;DB;2 GT:DP:HQ 0:0:48:1:51,51 1:0:48:0:51,51 1/1:48:1...
20 17330 T A 3 GQ RS=0;DP=1;AF=0.0;GT:DP:HQ 0:0:49:3:58,50 0:1:3:1:65,3 0/0:41:3
20 110906 rs604035 A G,T 67 PASS RS=0;DP=13;AF=0.333,0.667;AA=T;DB GT:DP:HQ 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 130237 T 47 PASS RS=0;DP=13;AA=T GT:DP:HQ 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 1234567 microsat1 CTC G,CTCT 50 PASS RS=0;DP=9;AA=G GT:DP: 0/1:38:4 0/2:17:2 1/1:40:3
```

CANCER

y_3

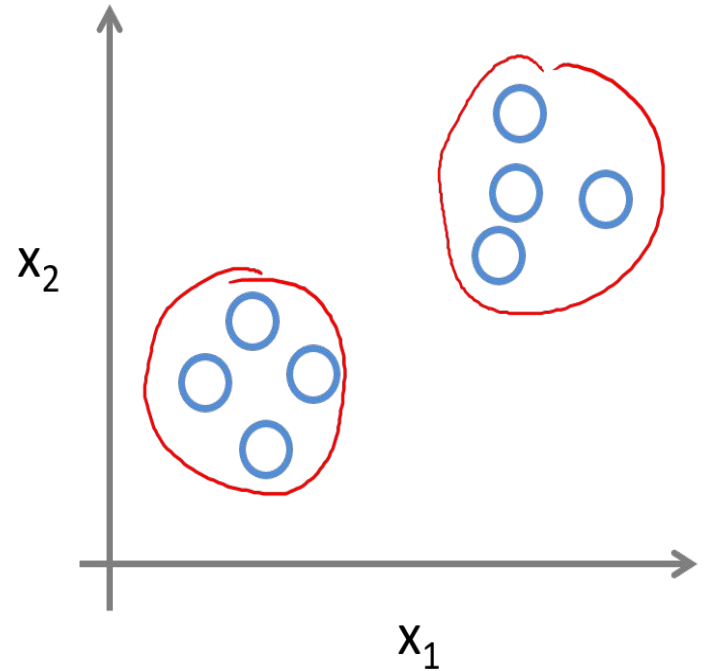
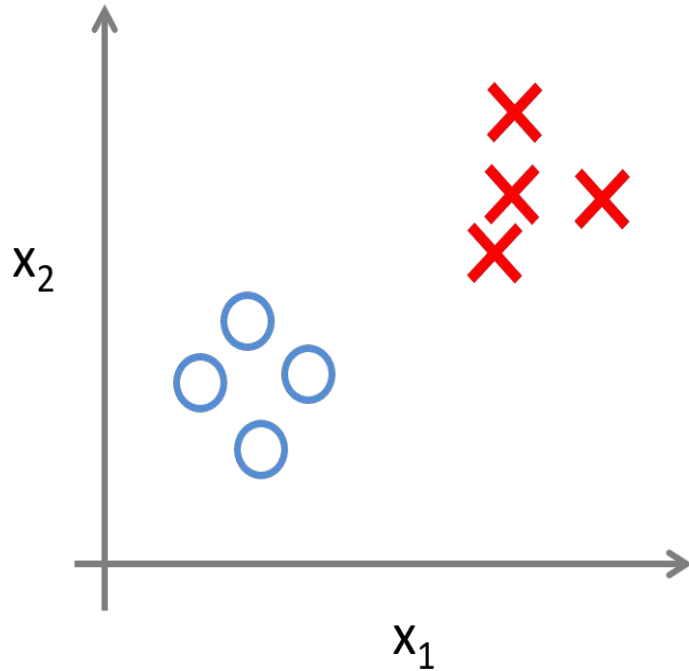
paciente03 X_3

Tipos de Aprendizaje: No supervisado



Cada instancia (sample) **no posee** etiqueta (izq). Los modelos a aplicar en estos casos buscan encontrar estructuras o grupos implícitas en los datos (ej. clusters)

Supervisado vs No supervisado



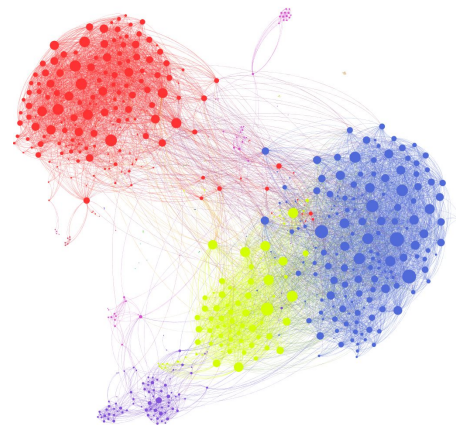
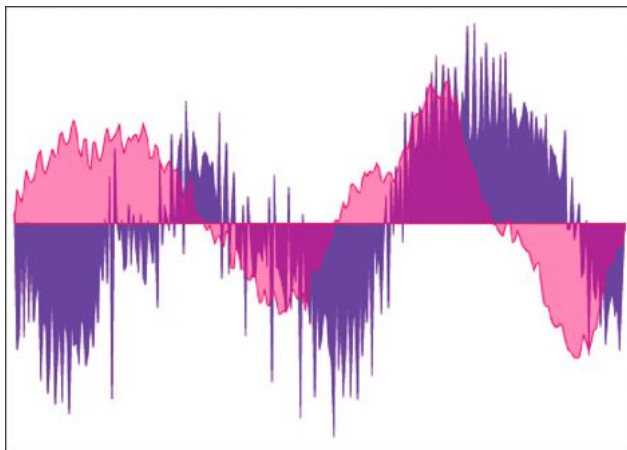
Tipos de Datos

- Estructurados / tabulares
- Imágenes
- Grafos-redes
- Lenguaje Natural (texto)
- Audio / Series de tiempo

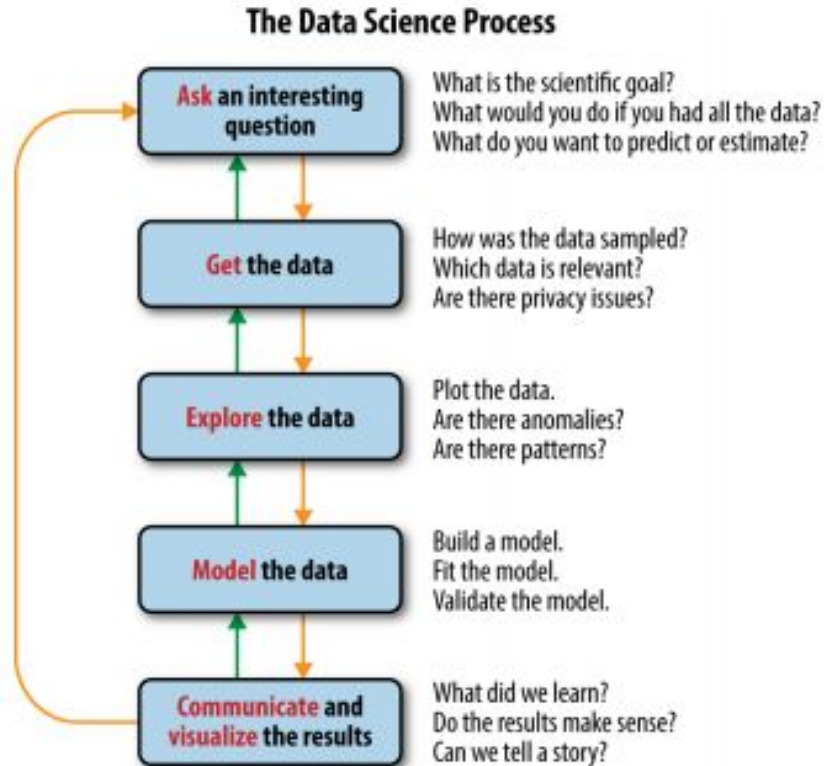
Formato de los datos

- .CSV
 - .xlsx
 - .txt
 - .tsv
 - .jpeg
 - SQL query

Tipos de Datos



Data Science Workflow



*Development Workflows for Data Scientists

1) Data Science Workflow: get the data

The Kaggle logo, featuring the word "kaggle" in a light blue, lowercase, sans-serif font.The Datos Argentina logo, featuring the text "Datos Argentina" in white on a blue background with a circular pattern. Below the text, it says "Portal de datos abiertos del Gobierno de la República Argentina. Acá encontrarás información pública, herramientas y recursos para desarrollar aplicaciones, visualizaciones y más."

Buenos Aires Data



Iniciativa de Datos Públicos y Transparencia de la Ciudad Autónoma de Buenos Aires.

Durante el curso trataremos de utilizar repositorios de datos abiertos, principalmente aquellos de la Ciudad de Buenos Aires, Provincia de Buenos Aires o Nación.

2) Data Science Workflow: Explore

Pre-processing

- Clean samples with NaNs
- Transform features
- Normalize data

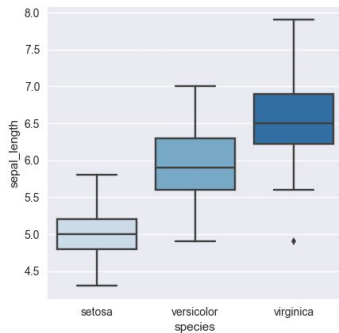
Exploratory Data Analysis

- Realizar estadísticas descriptivas
- Quitar outliers estadísticos
- Visualizar con Bar-plots, Box-plots, Scatter-plots, Count-plots

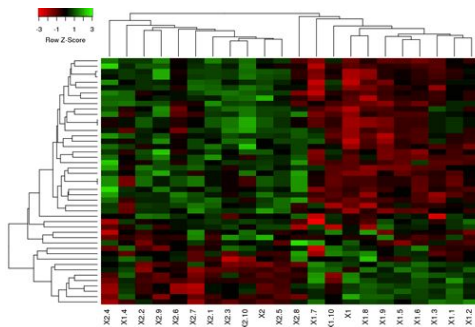
2) Exploratory Data Analysis

- Importar datos.
- Revisar si hay NaNs o valores faltantes.
- Filtrar los datos de interés.
- Transformar los datos (ej, tabla pivote)
- Computar estadísticas descriptivas (media, dev. std, percentiles)
- Medir correlación entre variables de interés
- Visualizar:

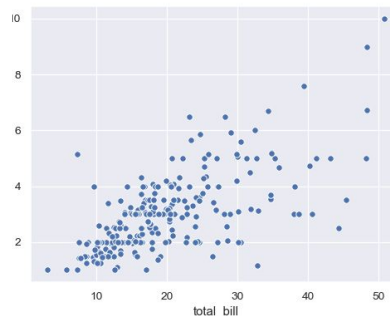
Boxplot



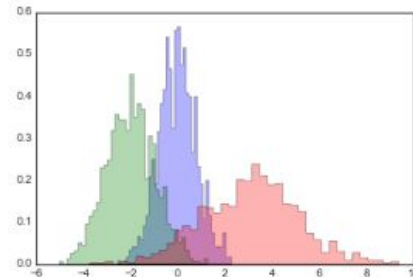
Heatmap



Scatter plot



Histogram



3) Data Science Workflow: Model

clasificación

regresión

Reducción de dimensionalidad

Detección de anomalías

clustering

Overview

Repositories

Projects

Stars

Followers

Following

Popular repositories

primeros_pasos_python

Este repositorio fue creado con el fin de ser utilizado en cursos donde se desea enseñar Python desde el comienzo y donde no existe el tiempo suficiente para dedicar a sus primeros pasos.

● Jupyter Notebook

★ 6

🔗 2

courses

Primeros Pasos Python

Fast.ai Courses

● Jupyter Notebook

ssh_localhost_port_tunneling_example

Forced from tunneling/ssh, localhost, port_tunneling_example

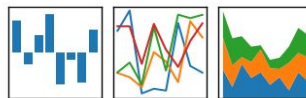
How to forward ports in ssh connection to connect to a jupyter notebook

martinpalazzo

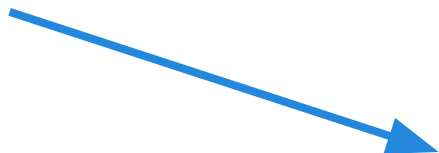
Tecnologías que utilizaremos



Librerías:



Python <> Anaconda <> Jupyter

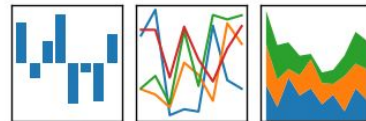


Numpy vs Pandas



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Calculo con matrices

- No admite nombres en cols
- No admite nombres en filas
- Diversidad en aplicaciones de cálculo
- Útil para lidiar con álgebra y operaciones matriciales

Atajos de Numpy acá:

https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Numpy_Python_Cheat_Sheet.pdf

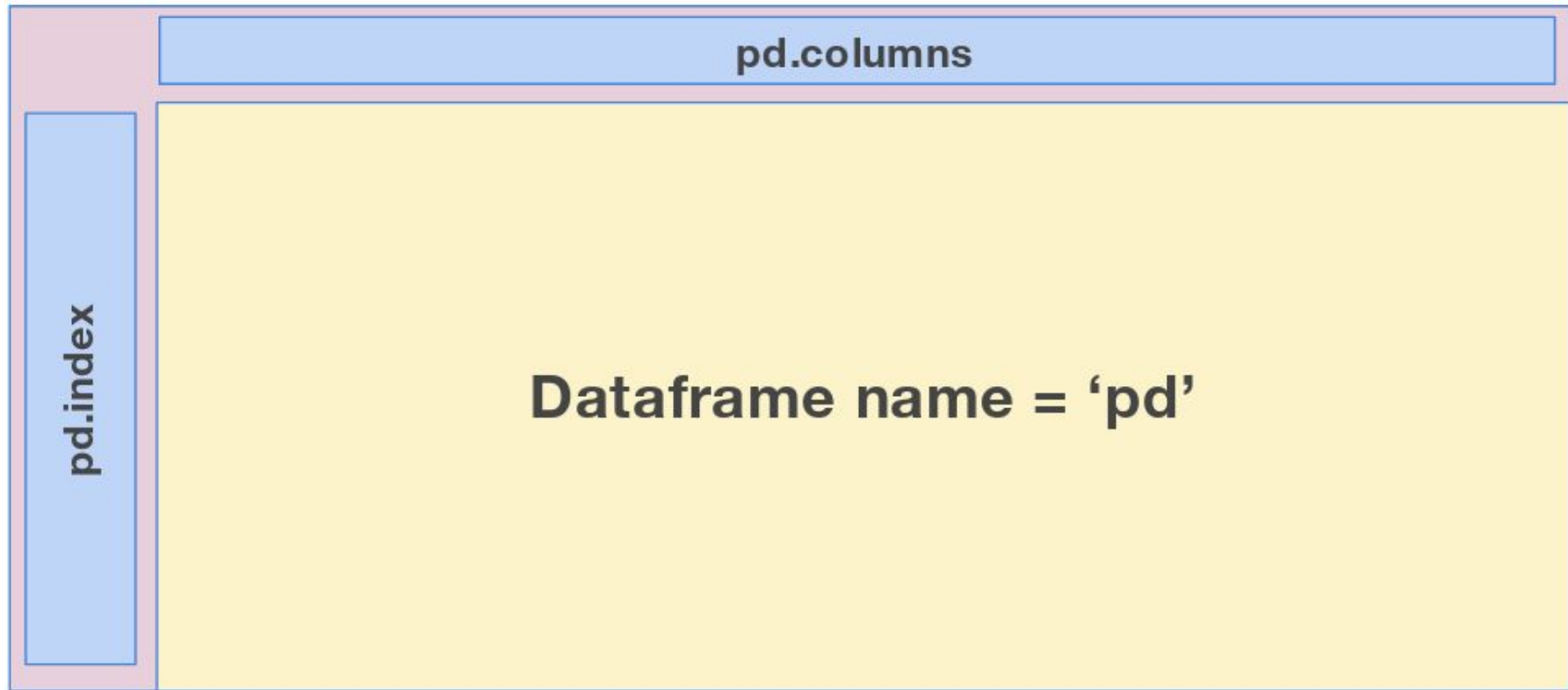
Gestor de datasets en Dataframes (DFs)

- Admite nombre de columnas
- Admite nombre de filas
- Diversas funciones sobre DFs.
- Útil para lidiar con datos, limpiar, pre procesar.

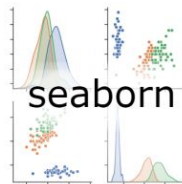
Atajos de Pandas acá:

https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

Pandas Dataframe



Visualización



Librerías de visualización de datos

- Matplotlib es la principal librería de visualización en Python.
- Seaborn corre sobre matplotlib y posee algunas mejoras de estética.
- Tipos de gráficos a realizar:
 - Countplot (graficos de barra)
 - Heatmap (mapas de calor)
 - Boxplot (diagrama de cajas y bigote)
 - Series de tiempo
 - Scatter plot (diagrama de puntos)
 - Distplot (distribuciones y densidades)

Atajos de Matplotlib acá:

https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Python_Matplotlib_Cheat_Sheet.pdf

Tipos de variables en Python



```
a = 3
b = 0.4
c = False
d = "Quiero analizar datos"
e = [2,3,4,5,6]
f = [[2,3,4],[1,0,40]]
```

a = Integer

b = Float

c = Boolean

d = String

e = Numpy Array (1,5)

f = Numpy Array (2,3)

A agarrar la PyLA



Preproc. + EDA con data de arboles




Import libraries



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Import Data



```
molinetes = pd.read_csv('/home/human/Dropbox/clustera/molinetes_historico.csv', delimiter=';', index_col=['PERIODO'])
```