

clusterAI 2020
ciencia de datos en ingeniería industrial
UTN BA
curso I5521

clase_00



clusterAI

Arte generativo



Mario Klingemann

agenda_clase00

- Consignas de la materia, presentación del equipo docente
- Features, samples, the curse of dimensionality
- Tipos de Aprendizaje, Supervisado vs No Supervisado
- Tipos de Datos
- Formato de los datos
- Data Science Workflow
- Primeras prácticas con Python

Propuesta de valor del curso

Preparar a los futuros ingenier@s, profesionales, entusiastas y emprendedores para lidiar con complejidad en el contexto de la 4ta revolución industrial*.

*<https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>

Misión del curso

Lograr que el estudiante termine el curso incorporando:

1. Conocimiento básico-intermedio de análisis de datos con python
2. Métodos clásicos y no tanto de machine learning.
3. Realización y comunicación de un proyecto con modelos y tecnicas de data science y machine learning sobre datos reales.
4. Incorporándose a una comunidad activa y creciente de ciencia de datos.

Herramientas del curso

- Comunicación semanal: por medio del google groups del curso.
- Comunicación diaria y comunitaria: canal de slack.
- Contexto del contenido: Diapositivas de clase (apuntes solamente de soporte) + jupyter notebooks + canal de youtube.
- Programar: Python (Anaconda, Jupyter) para programar.
- Teoría base: Libros.
- Ejercitación: Guía de ejercicios.

Mentores y colaboradores

Fernando Buttafuoco, Mauro Lucini, Angel Daniel Tabia, Gabriel Boso, Martin Moro, Santiago Chas, Lucas Chicco, Mireya mamani, Alejandro Bachur Solari, Ezequiel Vannucchi, Florencia Sanchez, Joaquin Magallanes, Conrado Ochoa, Gabriel Castaño, Agustín Carpaneto, Lautaro Rshaid, Santiago Rubio, Nicolás Alejandro Bogliolo, Francisco Chedufau.

docentes clusterAI



Nicolas Aguirre

Docente asistente ClusterAI
Machine Learning UTN FRBA &
Université de Technologie de Troyes
Doctorant UTN-UTT
Master OSS (UTN-UTT)
Ingeniero Industrial UTN BA



Agustin Velazquez

Docente asistente ClusterAI
Data Analytics Developer en
AlixPartners
Master OSS (UTN-UTT)
Docente Inv. Op. UTN BA
Ingeniero Industrial UTN BA



Martin Palazzo

Docente coordinador ClusterAI
Machine Learning Instituto IBioBa Max Planck
& Université de Technologie de Troyes
Doctorant UTT UTN
Docente Inv. Operativa. UTN BA
Master OSS (UTN-UTT)
Ingeniero Industrial UTN BA

Miembros Cluster



Matias Callara

Data Scientist Roche (CH)
Doctorado en Data Science (Francia)
Master OSS (UTN-UTT)
Ingeniero Industrial UTN BA



Sebastian Pinto

Docente UTN Cooperativismo
Consejero depto. Industrial
Master OSS (UTN-UTT)
Ingeniero Industrial UTN BA

Cluster online



clusteraigroup@gmail.com



facebook.com/clusteraigroup/



twitter.com/clusteraigroup



github.com/clusteraigroup



linkedin.com/company/clusteraigroup

Bibliografía

Recomendamos tener a mano los siguientes libros en orden de prioridad:

- Data Science Handbook (VanderPlas)
- Introduction to Statistical Learning (Tibshirani)
- Deep Learning Book (Goodfellow)
- Elements of Statistical Learning (Tibshirani)
- Hands on Machine Learning with Scikit Learn (Geron)
- Machine Learning & Pattern Recognition (Bishop)

Estructura del curso

Datos Estructurados

Análisis Exploratorio de Datos

Aprendizaje Supervisado: Clasificación y regresión.

Reducción de dim. + regularización

Aprendizaje No supervisado: clustering

Procesamiento del Lenguaje Natural (NLP)

Análisis del sentimiento

Word Embeddings

Deep Learning

Redes Neuronales: Clasificación

Redes Neuronales: Autoencoders

Redes Neuronales: Series de tiempo

Requisitos de aprobación

- *Asistencia y seguimiento de las clases (4 faltas máximo).*
- Entrega de trabajo práctico integral.
- Poster para ser presentado en el evento de fin de cursada.
- Aprobar parcial teórico-práctico.

Informacion



Que diferencias les llaman la atención entre los dos vehículos?

Variables aleatorias



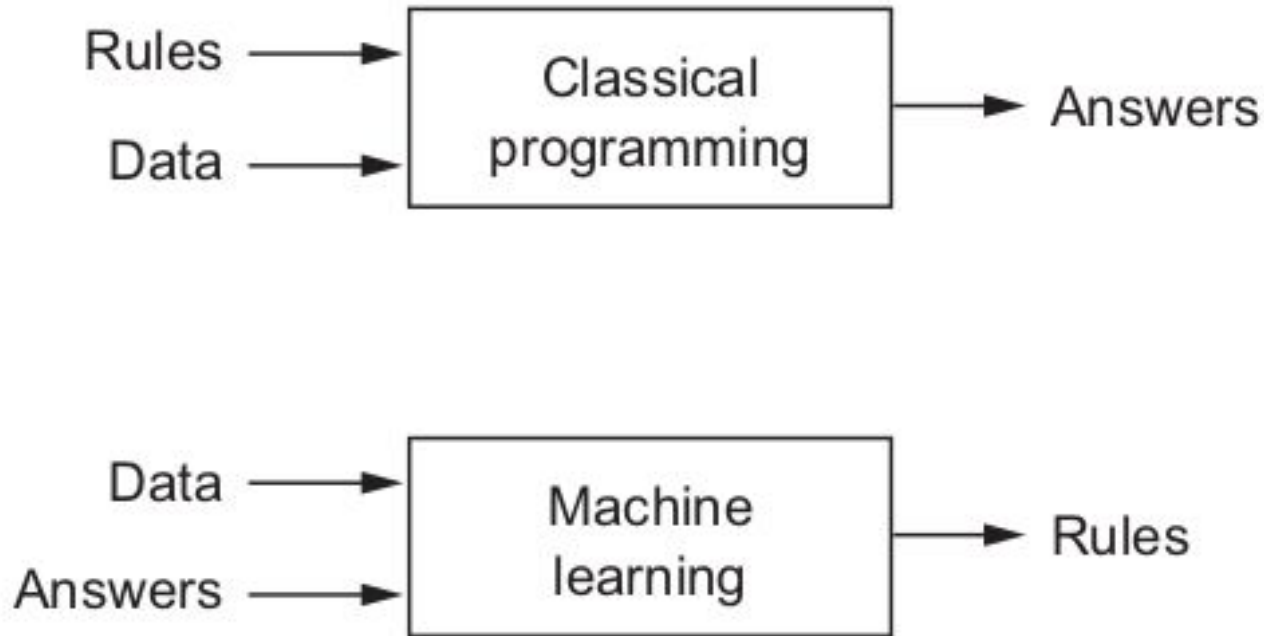
Cual es la probabilidad de encontrarte a alguien conocido en la calle? De que depende?



machine learning in a nutshell

**aprender y construir modelos
desde los datos.**

machine learning in a nutshell



Samples (instancia) & Features (atributos)

Los datos estarán caracterizados por dos indicadores: Samples & Features

Samples/instancias

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 ***n*** de samples.

$$S : \{x_1, x_2, x_3, \dots, x_n\}$$

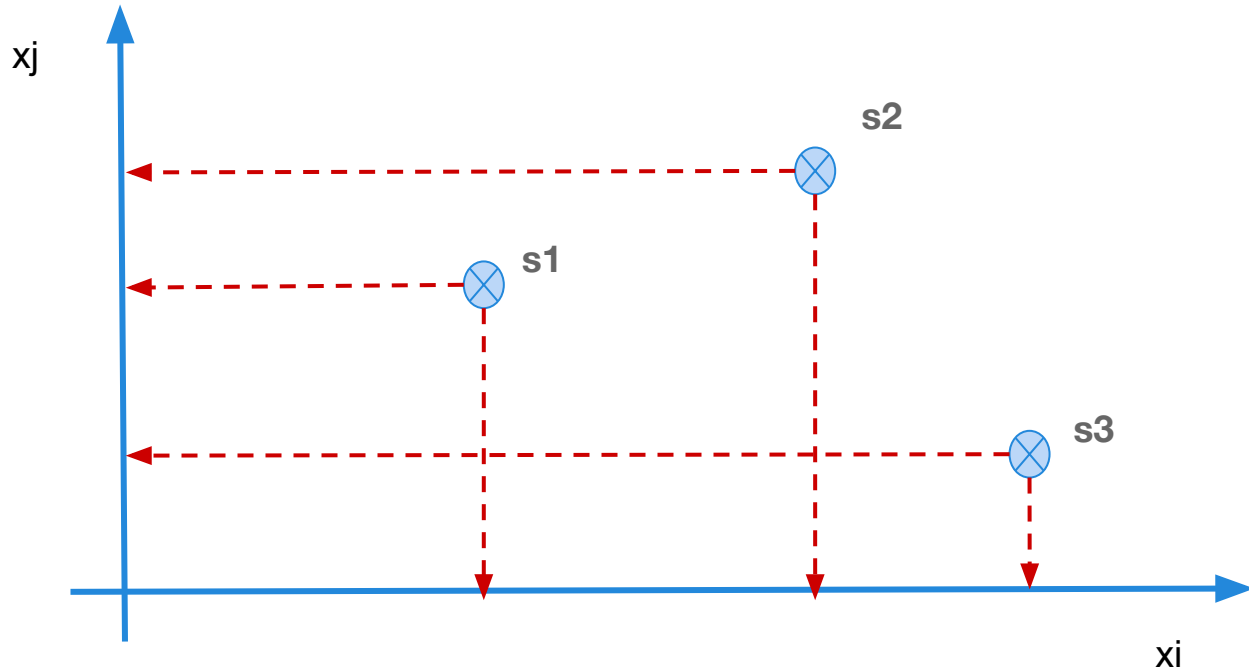
Samples & Features

$$x_i^T = [x_{i1}, x_{i2}, x_{i3}, \dots, x_{id},]$$

Features/variables/atributos/dimensiones

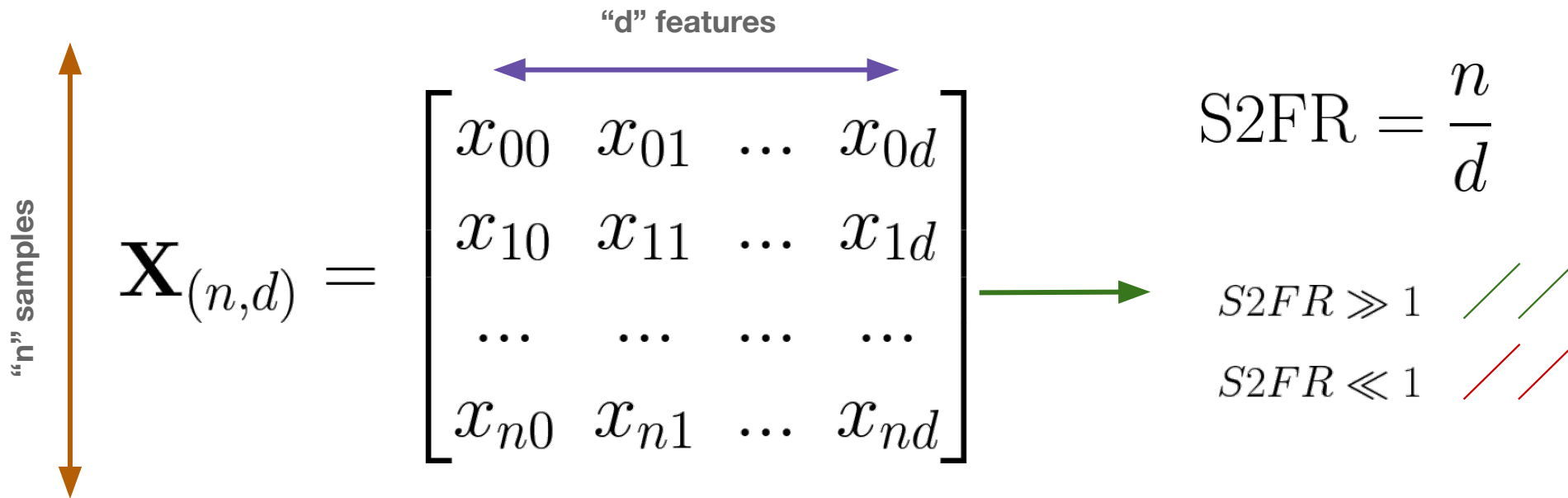
Denominamos features (atributos o mediciones) a las **variables** que definen y caracterizan a cada sample (instancia). La cantidad de features/variables que posea un sample es equivalente a la cantidad de **dimensiones** que describen a esa instancia en un espacio vectorial de ***d*** dimensiones. Entonces cada sample podemos considerarla un vector de ***d*** dimensiones. Nuestros datos “viven” en un espacio d-dimensional. Además, cada una de la dimensiones es considerada una variable aleatoria que sigue una distribución de probabilidad conocida o desconocida.

Samples (instancia) & Features (atributos)



¿Cuántas features/variables/dimensiones y cuántas instancias/samples hay en este ejemplo?

Sample-to-feature ratio


$$\mathbf{X}_{(n,d)} = \begin{bmatrix} x_{00} & x_{01} & \dots & x_{0d} \\ x_{10} & x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots & \dots \\ x_{n0} & x_{n1} & \dots & x_{nd} \end{bmatrix}$$
$$S2FR = \frac{n}{d}$$
$$S2FR \gg 1$$
$$S2FR \ll 1$$

Cuando el número n de samples < número de features d la relación entre instancias y dimensiones es menor a uno y eso implica mayor dificultad para poder explicar y describir el espacio donde viven nuestras instancias.

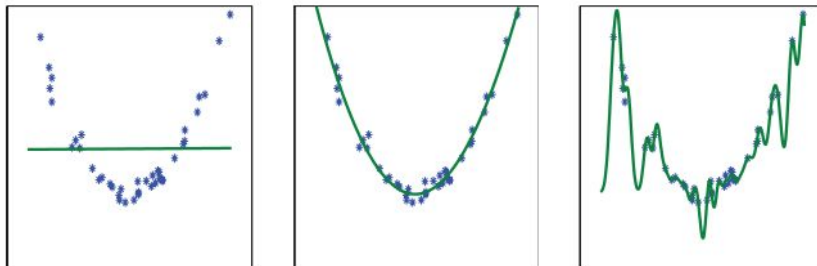
Aprender de datos

Nuestro problema tendrá una característica particular: no conocemos la distribución de densidad de probabilidad (PDF) de nuestros datos. Nuestros datos difícilmente coincidan con una distribución normal o una exponencial. Lo más probable es que pertenezcan a una distribución compleja en alta dimensión.

- Una opción es aproximarlos a distribuciones que ya conozcamos (metodo de maxima verosimilitud) aunque no necesariamente esa es la mejor opción.
- La alternativa de **Aprendizaje Estadístico / Machine Learning** es aproximar funciones específicas a nuestros datos de manera tal que puedan detectar patrones sin intentar determinar el tipo de distribución que explica los datos.

Aprender de datos

- Partiendo de un set de datos S aprenderemos una función “ $f(x)$ ” desconocida y será el estimador que utilizaremos.
- Nunca llegaremos a una “ $f(x)$ ” ideal que explique a la perfección nuestros datos, por ende tendremos cierto grado de error. La función $f(x)$ supone una distribución de probabilidad “ $p(x)$ ” que es incierta.
- Vamos a querer aprender una función que **generalice** bien para futuros datos nunca vistos. Es decir, una vez que encontramos el patrón en los datos disponibles de *entrenamiento*, esperamos que la $f(x)$ siga encontrando los mismos patrones para datos futuros nunca vistos.



Tipos de Aprendizaje Automático

Aprendizaje:

- **Supervisado**
- **No Supervisado**
- **Semi-supervisado**
- **Por refuerzo**

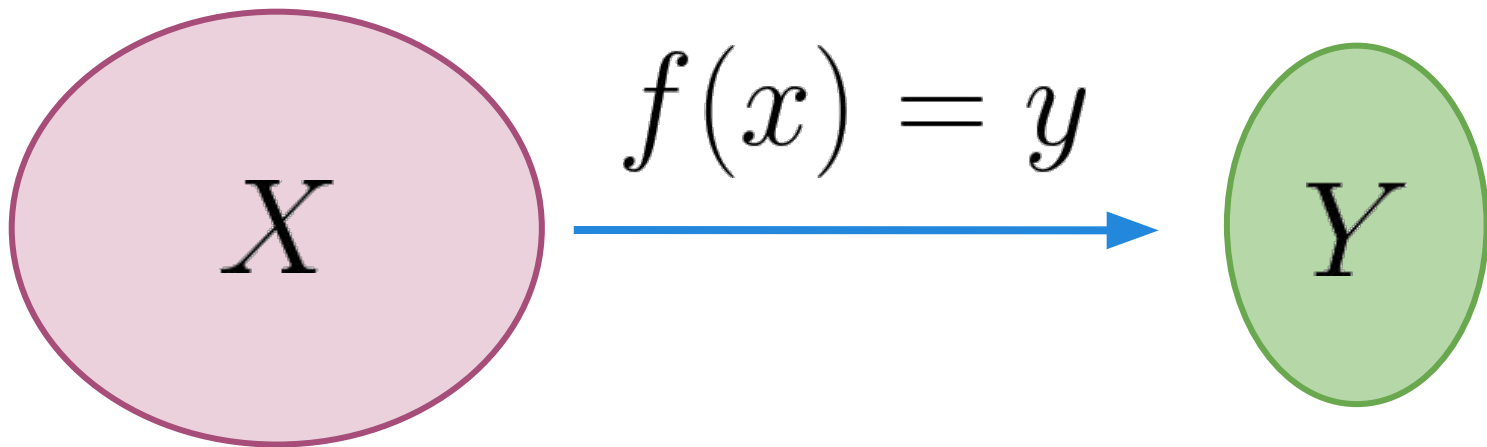
Aprendizaje Supervisado

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

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 .

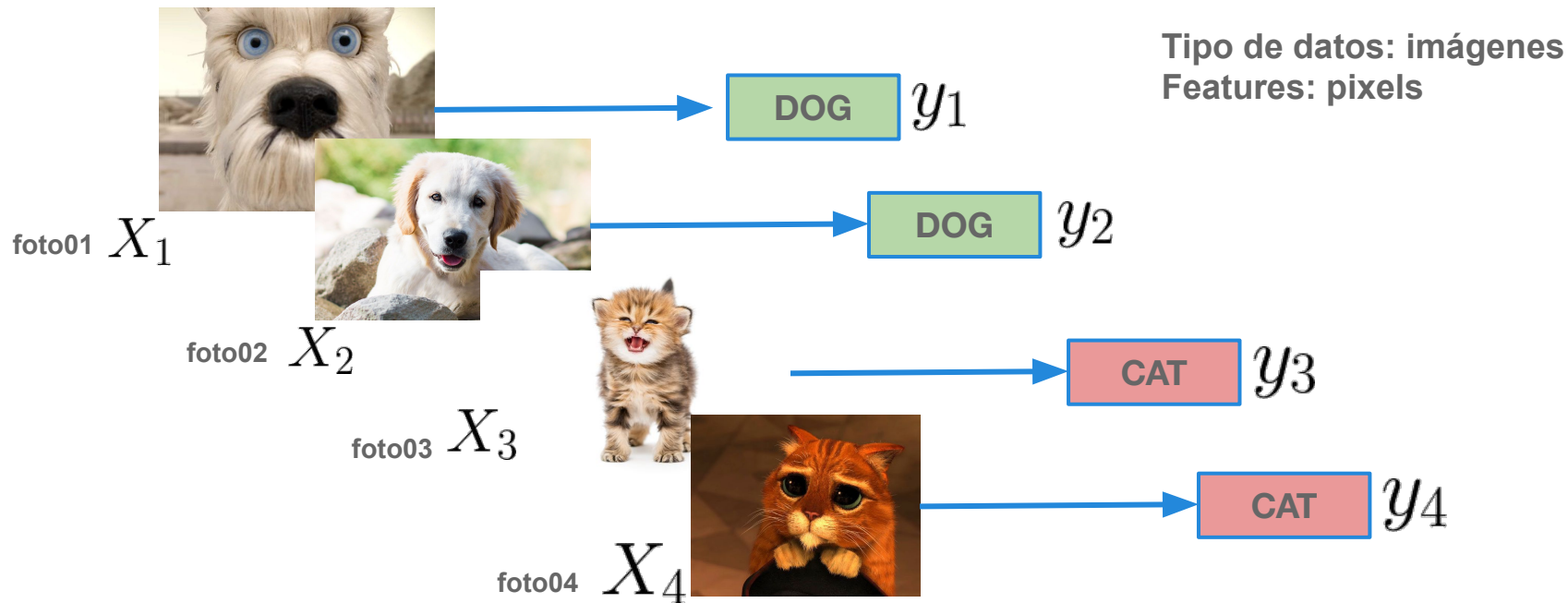
Solemos denominar a cada observación X_i como “sample” y a cada etiqueta Y_i como “label”.

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

paciente01 X_1

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=sequencingProgramV3.1
##reference=file:///seq/references/1000GenomePilot-NCBI36.fasta
##contig=ID=0,length=2438564,assembly=36,md5=f126cf86dc7317668f66b066,species="Homo sapiens",taxonomy>
##phasing=partial
##INFO=ID=0,Number=1,Type=Integer,Description="Number of Samples With Data"
##INFO=ID=DP,Number=1,Type=Integer,Description="Total Depth"
##INFO=ID=AF,Number=1,Type=Float,Description="Allele Frequency"
##INFO=ID=AA,Number=1,Type=String,Description="Ancestral Allele"
##INFO=ID=SB,Number=0,Type=Flag,Description="GBDP membership, build 129"
##INFO=ID=2,Number=0,Type=Flag,Description="HighP2 membership"
##FILTER=ID=Q1,Description="Quality below 10"
##FILTER=ID=Q10,Description="Less than 10% of samples have data"
##FORMAT=ID=GT,Number=1,Type=String,Description="Genotype"
##FORMAT=ID=DP,Number=1,Type=Integer,Description="Read Depth"
##FORMAT=ID=SB,Number=2,Type=Flag,Description="Sample type Quality"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAK0001 SAK0002 SAK0003
20 14370 rs604207 C A 20 PASS RS-Q:DP=1;AF=0.0;DB;2 GT:DP:SB 0:0:48:1:51,51 1:0:48:0:51,51 1/1:48:1...
20 17330 T A 3 GQ RS-Q:DP=1;AF=0.0:17 GT:DP:SB 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-Q:DP=13;AF=333,0.667;AA=T;DB GT:DP:SB 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 1302237 T 47 PASS RS-Q:DP=13;AA=T GT:DP:SB 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-Q:DP=9;AA=G GT:DP:SB 0/1:36:4 0/2:17:2 1/1:40:3
```

CANCER y_2

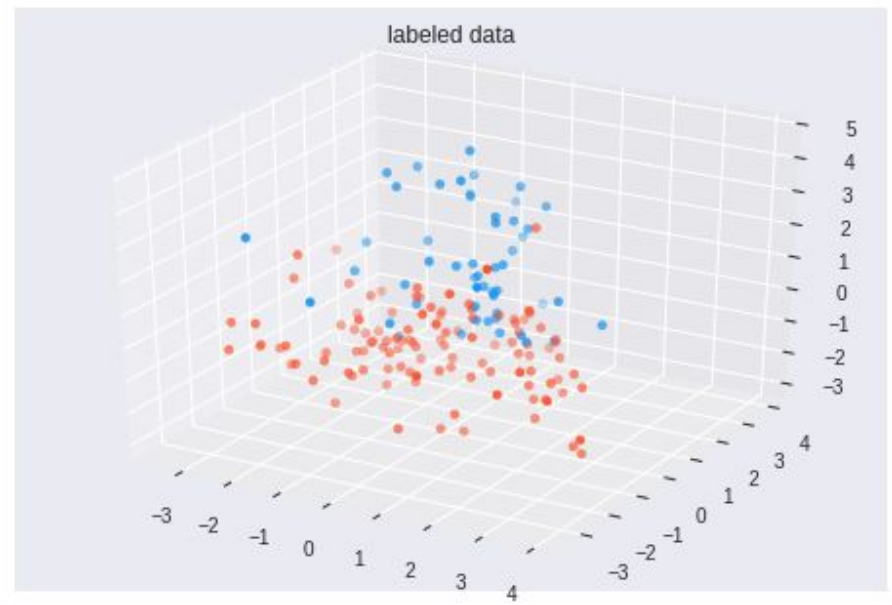
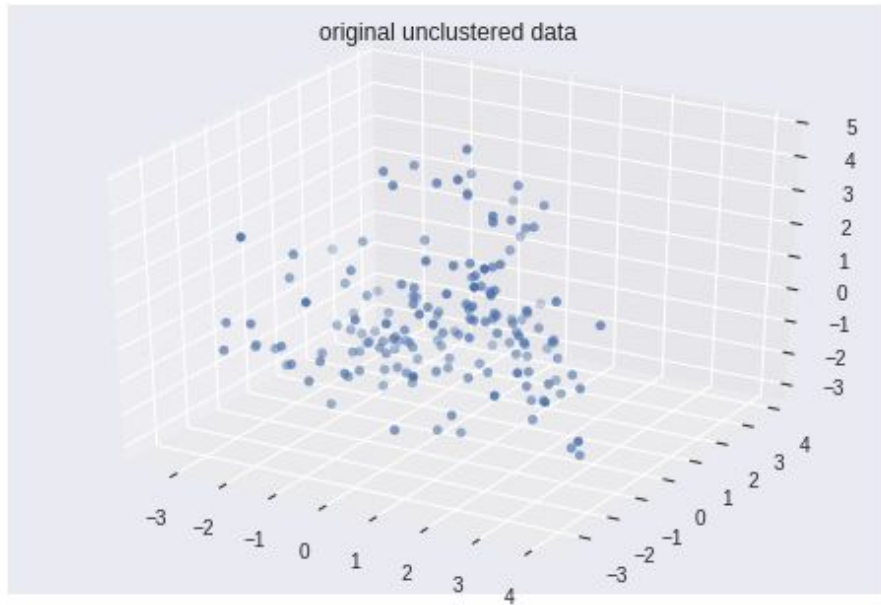
paciente02 X_2

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=sequencingProgramV3.1
##reference=file:///seq/references/1000GenomePilot-NCBI36.fasta
##contig=ID=0,length=2438564,assembly=36,md5=f126cf86dc7317668f66b066,species="Homo sapiens",taxonomy>
##phasing=partial
##INFO=ID=0,Number=1,Type=Integer,Description="Number of Samples With Data"
##INFO=ID=DP,Number=1,Type=Integer,Description="Total Depth"
##INFO=ID=AF,Number=1,Type=Float,Description="Allele Frequency"
##INFO=ID=AA,Number=1,Type=String,Description="Ancestral Allele"
##INFO=ID=SB,Number=0,Type=Flag,Description="GBDP membership, build 129"
##INFO=ID=2,Number=0,Type=Flag,Description="HighP2 membership"
##FILTER=ID=Q1,Description="Quality below 10"
##FILTER=ID=Q10,Description="Less than 10% of samples have data"
##FORMAT=ID=GT,Number=1,Type=String,Description="Genotype"
##FORMAT=ID=DP,Number=1,Type=Integer,Description="Read Depth"
##FORMAT=ID=SB,Number=2,Type=Flag,Description="Sample type Quality"
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAK0001 SAK0002 SAK0003
20 14370 rs604207 C A 20 PASS RS-Q:DP=1;AF=0.0;DB;2 GT:DP:SB 0:0:48:1:51,51 1:0:48:0:51,51 1/1:48:1...
20 17330 T A 3 GQ RS-Q:DP=1;AF=0.0:17 GT:DP:SB 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-Q:DP=13;AF=333,0.667;AA=T;DB GT:DP:SB 0:0:54:7:56,60 0:0:48:4:51,51 0/0:41:2
20 1302237 T 47 PASS RS-Q:DP=13;AA=T GT:DP:SB 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-Q:DP=9;AA=G GT:DP:SB 0/1:36:4 0/2:17:2 1/1:40:3
```

CANCER y_3

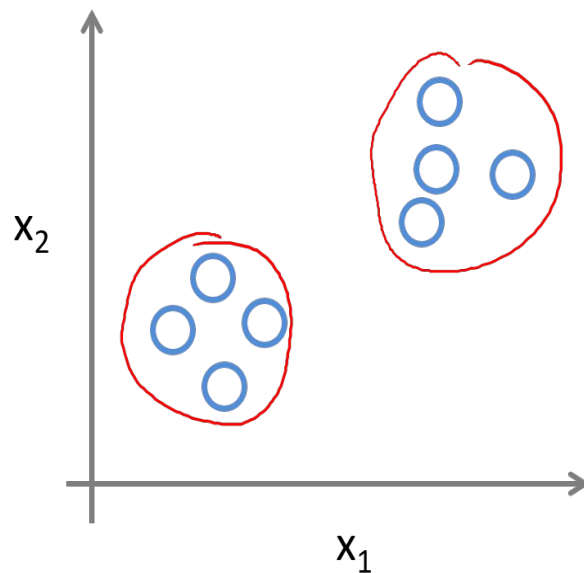
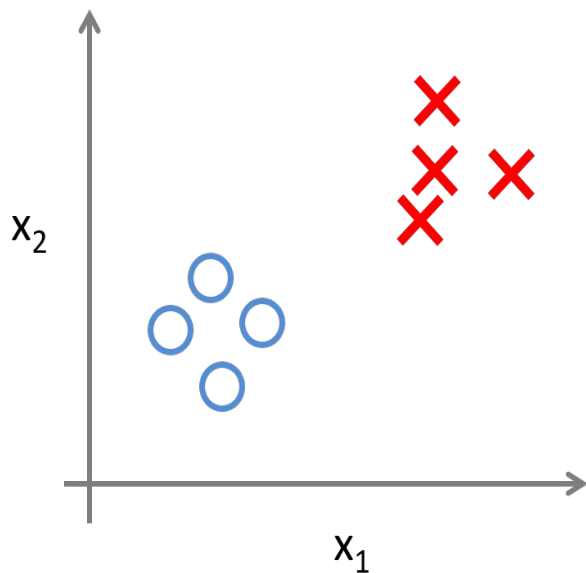
paciente03 X_3

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



Izquierda: Datos etiquetados. Derecha: Datos sin etiquetar estructurados en clusters.

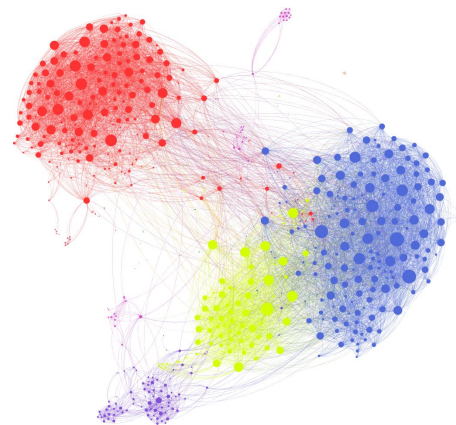
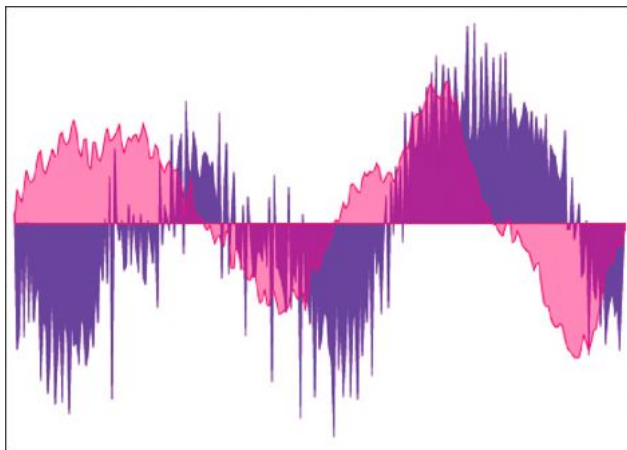
Tipos de Datos

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

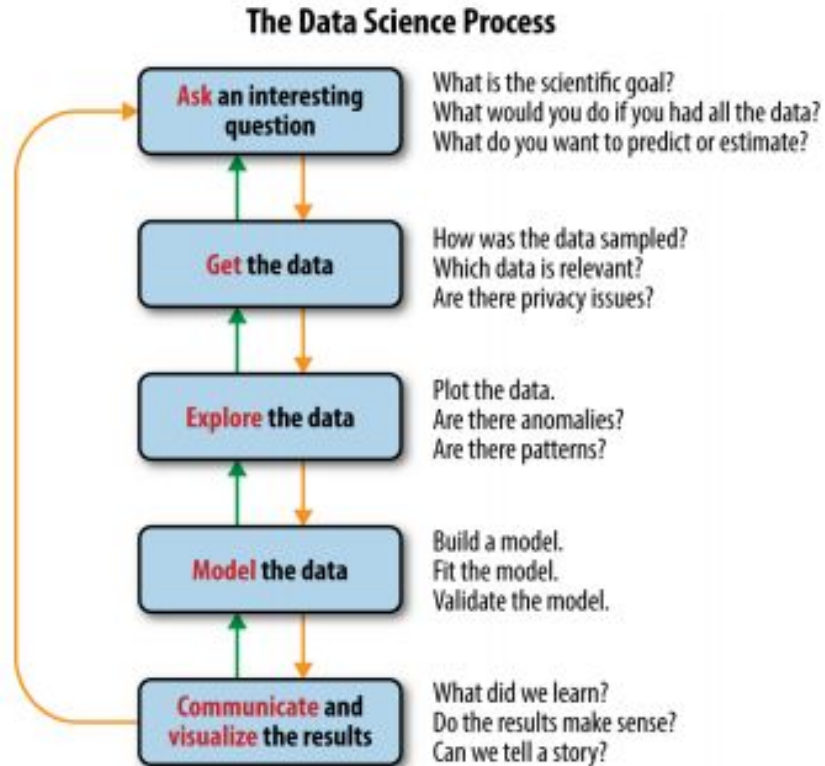
Formato de los datos

- .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

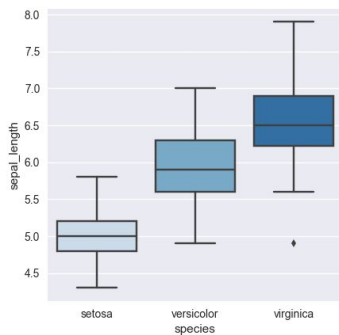
Exploratory Data Analysis

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

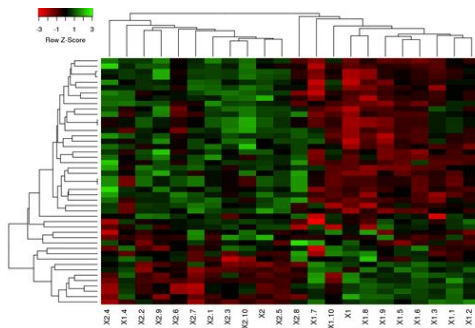
2) Exploratory Data Analysis (EDA)

- 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:

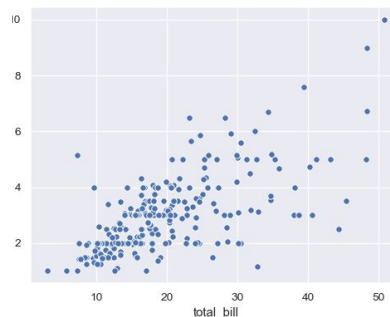
Boxplot



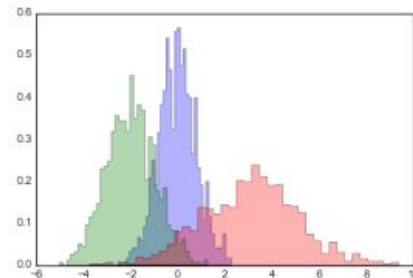
Heatmap



Scatter plot



Histogram



3) Data Science Workflow: Model

clasificación

regresión

Reducción de dimensionalidad

Detección de anomalías

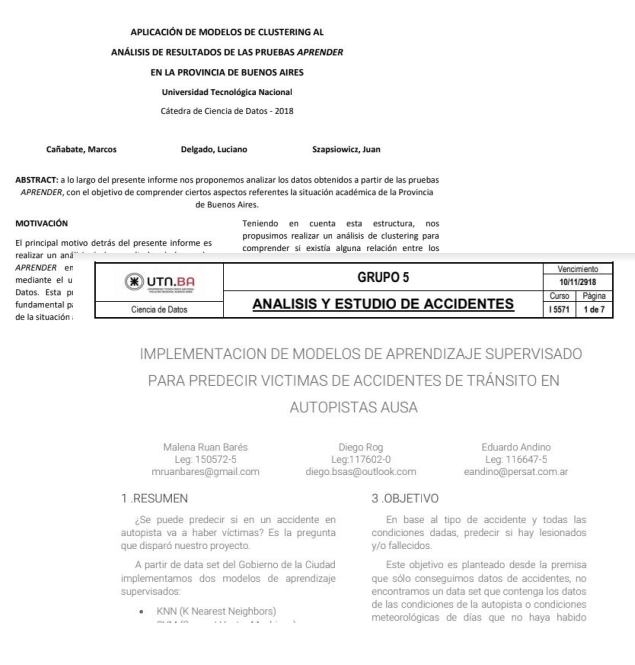
clustering

4) Communicate Results

Poster explicando el proyecto



Reporte técnico del proyecto



Exposición de posters



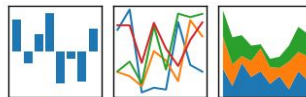
Repositorio en github del proyecto



Tecnologías que utilizaremos



Librerías:



Python <> Anaconda <> Jupyter

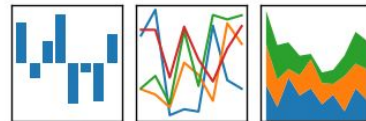


Numpy vs Pandas



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

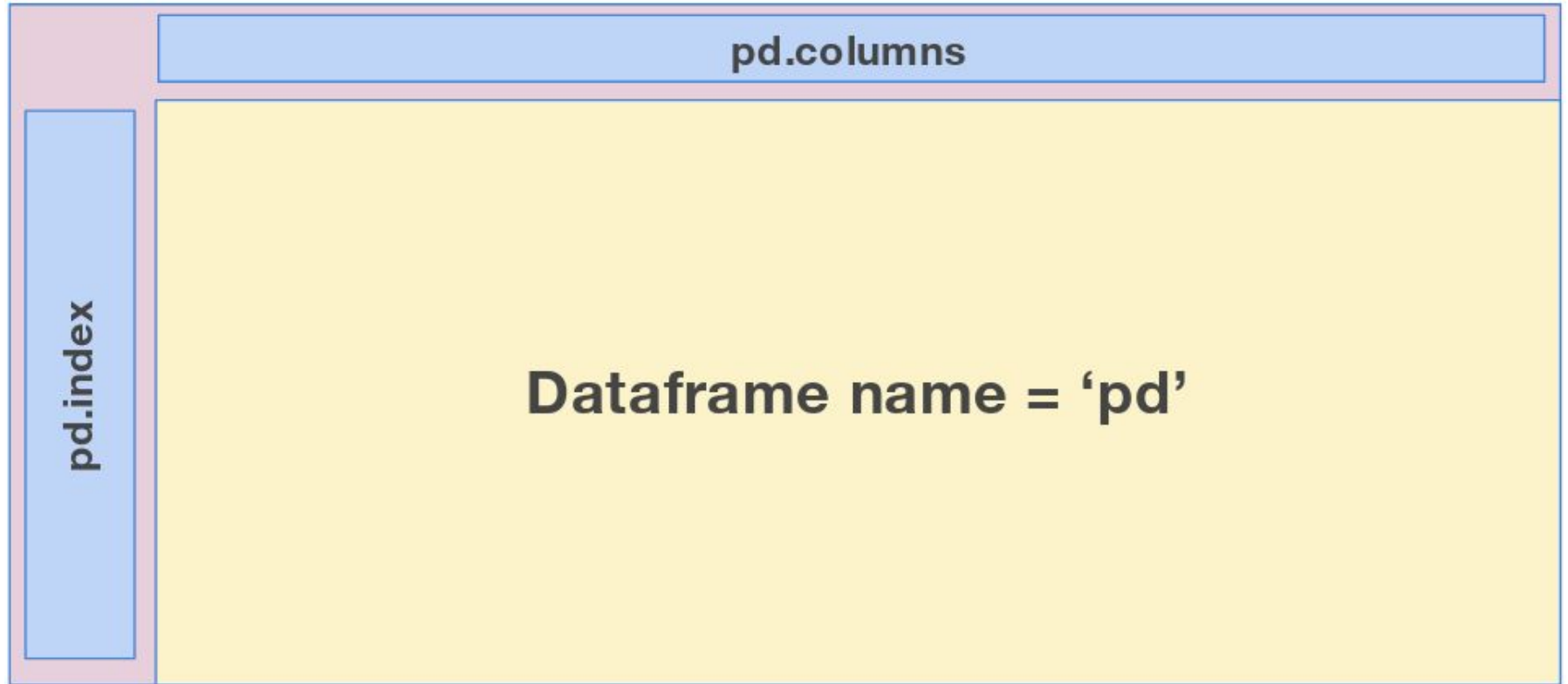
Gestor de datasets en Dataframes (DFs)

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

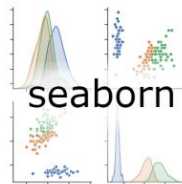
Atajos de Pandas acá:

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

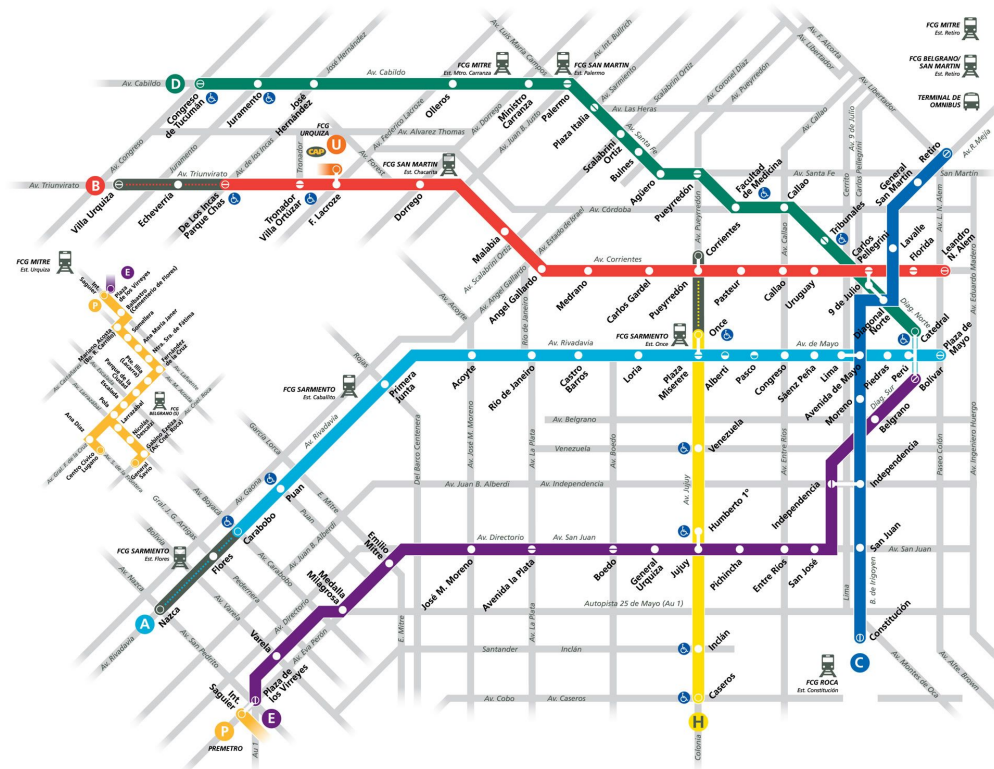
e = Numpy Array (1,5)

f = Numpy Array (2,3)

A agarrar la PyLA



Preproc. + EDA con data de subtes




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'])
```

Exploratory Data Analysis

