

“Introducción”

APRENDIZAJE AUTOMÁTICO

PROGRAMA DE CIENCIA DE DATOS

Profesor: MSc. Felipe Meza



July 15, 2019

Agenda

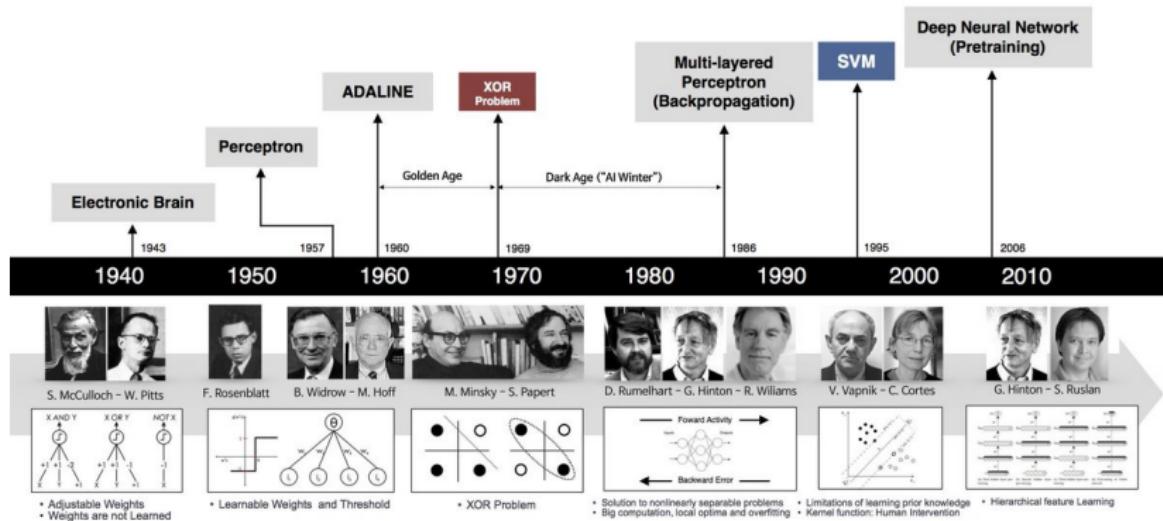
- ① Introducción al Aprendizaje Automático.
- ② Estadística? ML? Patrones?
- ③ ML y Programación Tradicional.
- ④ Etapas básicas de un sistema de ML.
- ⑤ Tipos de Aprendizaje.
- ⑥ Tareas del Aprendizaje.
- ⑦ Instancias, Atributos y Clases.
- ⑧ Metodología de Diseño.
- ⑨ Métricas para evaluar lo aprendido.
- ⑩ AMBIENTE PYTHON.
- ⑪ Pre-procesado.

Introducción al Aprendizaje Automático

- Nos encontramos en la era de los datos:
 - Todo queda registrado, tanto en el ámbito comercial como científico.
 - Volumen de datos crece, pero la comprensión de los mismos?
 - Se estima que el volumen de datos se duplica cada 20 meses.
- La extracción de información de valor (científico o comercial) representa una gran oportunidad.
- Dicha extracción se lleva a cabo mediante el descubrimiento de patrones en los datos a.k.a. *patrones estructurales*.
- El trabajo del *científico de datos* consiste en usar técnicas para extraer dichos patrones y usar los resultados para predecir nuevos eventos.

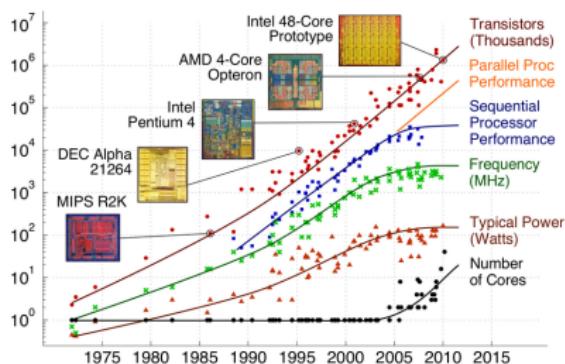
Introducción al Aprendizaje Automático

- Un recorrido por la historia de AI/ML:

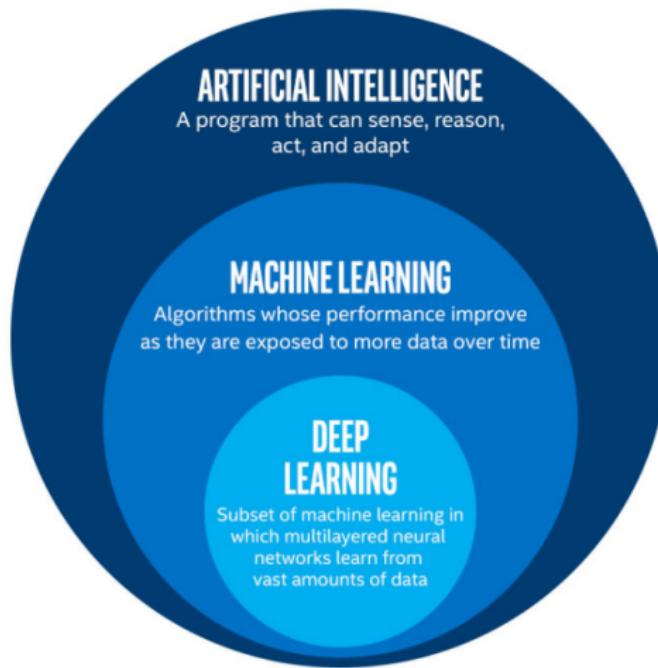


Introducción al Aprendizaje Automático

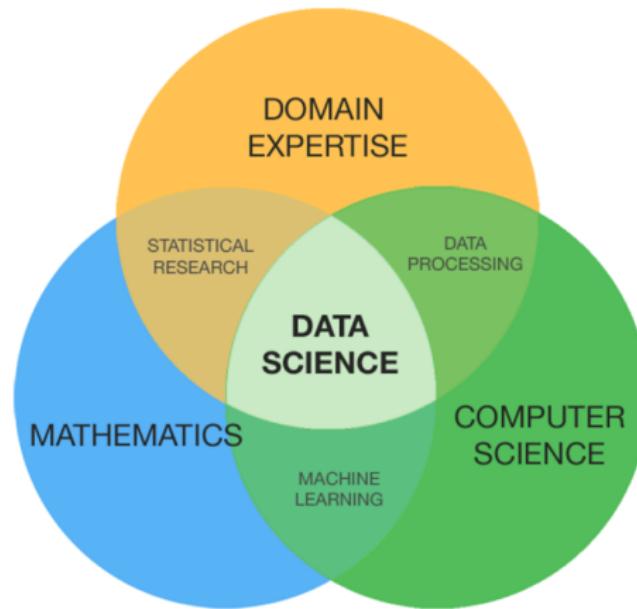
- Factores que hacen HOY posible AI/ML:
 - Capacidad computacional.
 - Generación actual de datos.



Introducción al Aprendizaje Automático



Introducción al Aprendizaje Automático



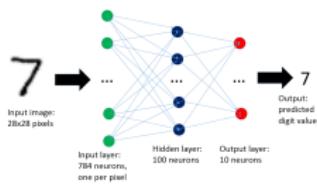
Introducción al Aprendizaje Automático

• MACHINE LEARNING (ML) :

- Conjunto de mecanismos que permiten convertir los datos en información/conocimiento útil (Peter Harrington).
- Construcción de programas computacionales que mejoren automáticamente con la experiencia (Tom Mitchell).
- Máquinas capaces de hacer tareas sin haber sido explícitamente programadas para eso (Arthur Samuel).
- También se le conoce como aprendizaje automático y utiliza técnicas de los campos de computación, matemática, entre otros.

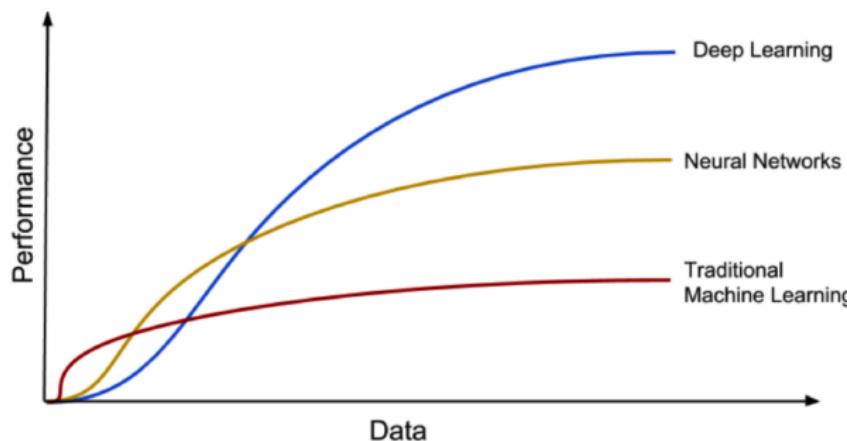


Self Driving Cars

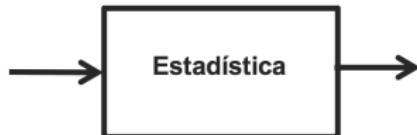


- **DEEP LEARNING (DL) :**

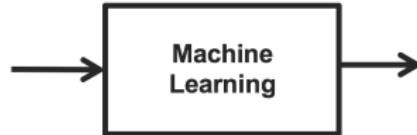
- Sub-campo de ML y puede verse como sinónimo de ANN's con múltiples capas y grandes cantidades de datos.
- *Las representaciones son jerárquicas !!!*



Estadística? Machine Learning? Patrones?

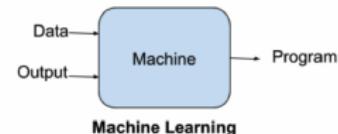
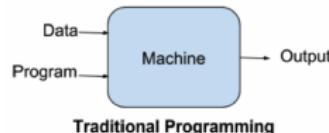
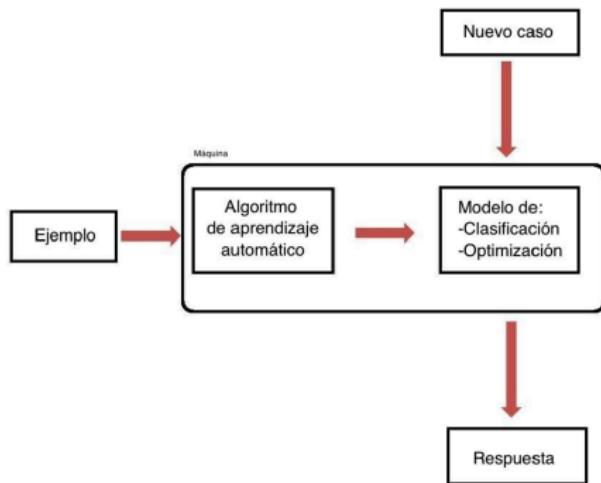


- Se busca la relación entre variables.
- Las ecuaciones matemáticas constituyen una herramienta vital.
- Campo de las matemáticas.
- Estimar, Hipótesis, Data Points, Respuestas.
- **INFERENCIA**



- Se busca aprender de los datos.
- Los algoritmos de aprendizaje constituyen una herramienta vital.
- Campo de CS (AI).
- Aprender, Clasificaciones/Regresiones, Instancias, Etiquetas.
- **GENERALIZACIÓN**

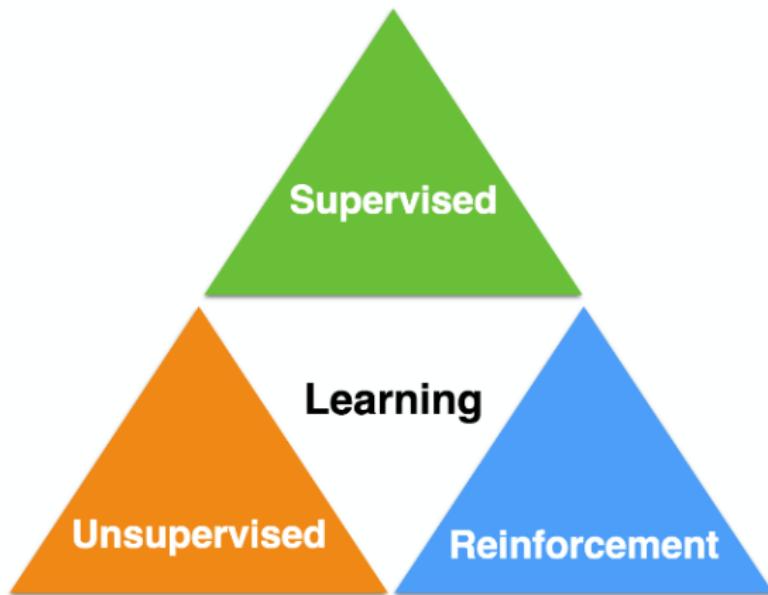
Aprendizaje Automático y Programación Tradicional



Etapas básicas de un sistema de ML



Tipos de Aprendizaje



Tipos de Aprendizaje

- **SUPERVISADO**

- Datos están etiquetados.
- Predice una salida futura.

- **NO SUPERVISADO**

- Datos NO están etiquetados.
- Encuentra una estructura “oculta”.

- **REFUERZO**

- Proceso de decisión.
- Sistema de recompensas.
- Aprende de una serie de acciones.

- **SUPERVISADO**

- Conjunto de datos etiquetados de la forma: $\{x_i, y_i\}_{i=1}^N$
- Donde x_i corresponde al vector de características con dimensiones $j = 1, \dots, D$, por lo tanto, cada característica se expresa como $x^{(j)}$
- Cada y_i se denomina etiqueta y puede consistir en una serie de clases de la forma $1, 2, \dots, C$

- **NO SUPERVISADO**

- Conjunto de datos NO etiquetados de la forma: $\{x_i\}_{i=1}^N$

- **REFUERZO**

- El algoritmo se desenvuelve en un **ambiente** como un vector de características, el algoritmo lleva a cabo **acciones** en cada **estado**, algunas acciones serán **recomendadas** otras **castigadas** en cada estado, para así pasar al siguiente, hasta aprender una **política**.

Tipos de Aprendizaje (**SUP** ó **NS?**)

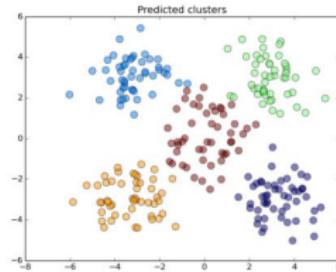
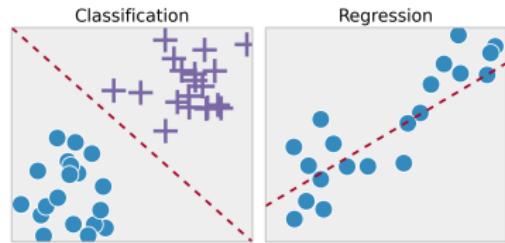
- A partir de un grupo de imágenes etiquetadas, reconocer a alguna persona.
- Agrupar a los estudiantes del curso por “estilos de aprendizaje”.
- A partir de los gustos de una persona y dado un conjunto de features (ritmo, genero musical etc.) recomendar una nueva canción.
- Analizar un conjunto de datos bancarios, y encontrar las “transacciones sospechosas” y etiquetar las fraudulentas.

Tipos de Aprendizaje (**SUP** ó **NS?**)

- A partir de un grupo de imágenes etiquetadas, reconocer a alguna persona. **SUP**
- Agrupar a los estudiantes del curso por “estilos de aprendizaje”. **NS**
- A partir de los gustos de una persona y dado un conjunto de features (ritmo, genero musical etc.) recomendar una nueva canción. **SUP**
- Analizar un conjunto de datos bancarios, y encontrar las “transacciones sospechosas” y etiquetar las fraudulentas. **NS**

Tareas del Aprendizaje

- Aprendizaje Supervisado
 - Clasificación
 - Predicción (Regresión)
- Aprendizaje NO Supervisado
 - Agrupamiento (Clustering)



Instancias, Atributos y Clases

Iris Data					
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
3	4.7	3.2	1.3	0.2	Iris setosa
4	4.6	3.1	1.5	0.2	Iris setosa
5	5.0	3.6	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
53	6.9	3.1	4.9	1.5	Iris versicolor
54	5.5	2.3	4.0	1.3	Iris versicolor
55	6.5	2.8	4.6	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
103	7.1	3.0	5.9	2.1	Iris virginica
104	6.3	2.9	5.6	1.8	Iris virginica
105	6.5	3.0	5.8	2.2	Iris virginica
...					

INSTANCES
samples/observations

ATTRIBUTES
features

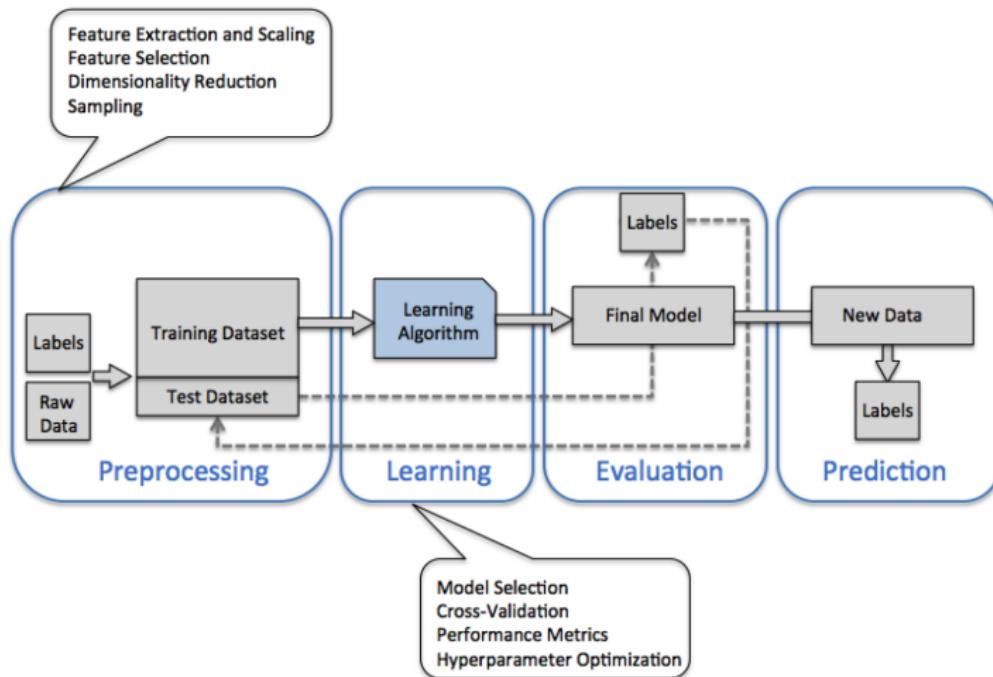
CLASSES
targets

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\mathbf{y} = [y_1, y_2, y_3, \dots, y_N]$$



Metodología de Diseño

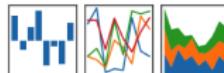


Métricas para evaluar el aprendizaje

- Exactitud de la clasificación.
- Pérdidas logarítmicas.
- Curva ROC.
- Matriz de confusión.
- MSE.
- R^2 .

AMBIENTE PYTHON

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



CONDA



python™



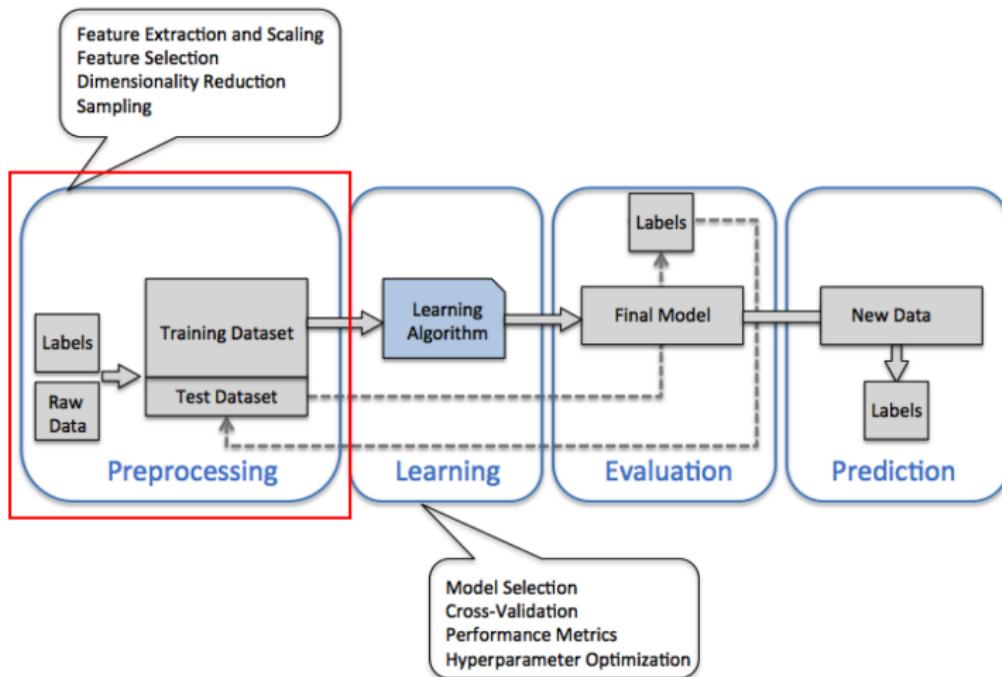
PYTORCH

Pre-procesado

Agenda

- ① Metodología de Diseño.
- ② Pre-procesado.
- ③ Preparación de los datos.
- ④ Algunas Tareas del Pre-procesado.
- ⑤ Análisis exploratorio de los datos (EDA).
- ⑥ Valores faltantes.
- ⑦ Outliers.
- ⑧ Datos no-balanceados.
- ⑨ Transformación de datos.
- ⑩ Reducción de dimensiones.

Metodología de Diseño



Pre-procesado

- Consiste en identificar partes o componentes del conjunto de datos que sean **incompletas**, imprecisas, incorrectas o irrelevantes, de manera tal que puedan ser **reemplazadas**, modificadas o removidas.
- Puede implicar también la **transformación** de los datos “puros” a otros formatos que faciliten su manejo por parte de los algoritmos de minería de datos.
- Incluye también la **reducción** de los datos a menores dimensiones para agilizar su procesamiento.
- En inglés varios **términos** se refieren a tales tareas: data preparation, cleaning, pre-processing, cleansing, wrangling.

Preparación de los datos

- En las metodologías de diseño generalmente en las **primeras etapas**, corresponde llevar a cabo las tareas de selección de datos, pre-procesado o transformación.
- En python, se recurre al uso de **librerías** como pandas que resultan ser muy buenas para las tareas asociadas a la preparación de los datos.
- Las labores de preparación no son un componente integral de los algoritmos de aprendizaje, sin embargo puede tomar un **tiempo** considerable dependiendo del conjunto de datos a analizar (80%-90% del proceso), por lo que se debe prestar especial atención.

Algunas Tareas del Pre-procesado

- Estandarizar, Normalizar.
- Análisis exploratorio de los datos.
- Valores faltantes.
- Outliers.
- Datos no-balanceados.
- Transformación de datos.

Normalizar, Estandarizar

- **Normalizar** (1) es llevar los datos a una nueva escala en un rango entre 0 y 1. Recomendado en casos donde los datos tengan múltiples escalas y donde los algoritmos sean sensibles a la escala.
- **Estandarizar** (2) consiste en llevar la distribución de los datos a una media de 0 y una desviación estándar de 1. Recomendado en casos donde el algoritmo es sensible a una distribución normal.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

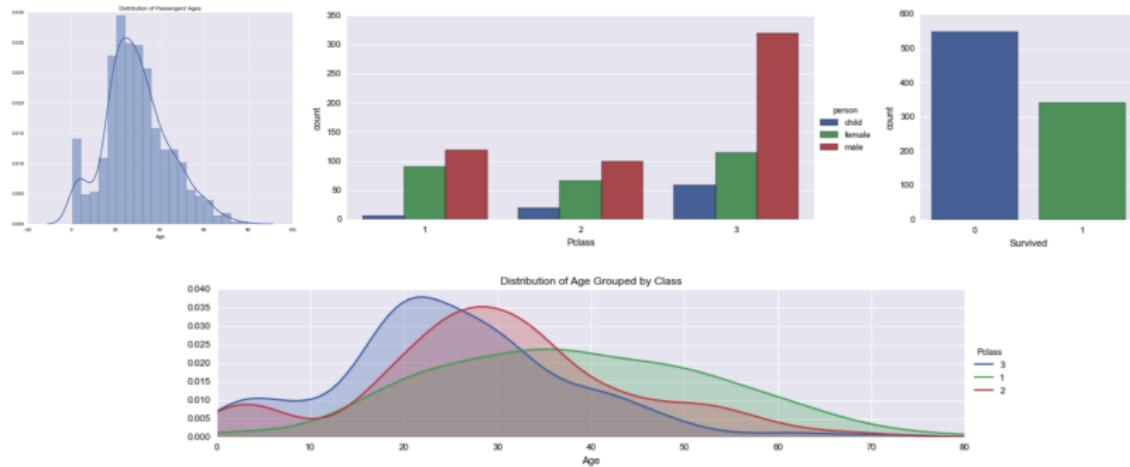
$$z = \frac{x - \mu}{\sigma} \quad (2)$$

Análisis exploratorio de los datos (EDA)

- Es la práctica del uso de métodos cuantitativos y **visuales** para comprender mejor un conjunto de datos sin tener que asumir hechos.
- Arrojar el conjunto de datos a un algoritmo y esperar los mejores resultados, **NO** es la mejor estrategia.
- Usualmente se lleva cabo una o varias de las siguientes actividades:
 - **Visualización** de un resumen estadístico del conjunto de datos.
 - **Exploración** visual de cualquier relación que pueda tener cada atributo con la clase que nos interesa predecir.
 - Mediante diagramas de dispersión **observar** cualquier tipo de agrupamiento que se pueda presentar en los datos.

Análisis exploratorio de los datos (EDA)

Ejemplo de EDA con conjunto de datos TITANIC



Valores faltantes

- No existe un **método universal** para instancias con valores faltantes.
- Algunas técnicas **comunes** son:
 - Eliminar instancias.
 - Eliminar atributos.
 - Calcular “media” del atributo faltante.
 - Calcular “mediana” del atributo faltante.
 - Calcular “moda” del atributo faltante.
 - Usar regresión para estimar el valor del atributo faltante.



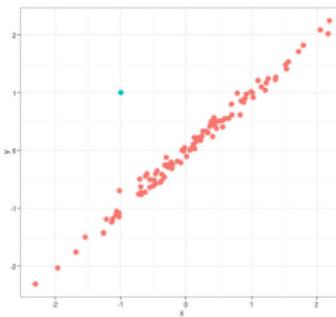
Valores faltantes

pandas

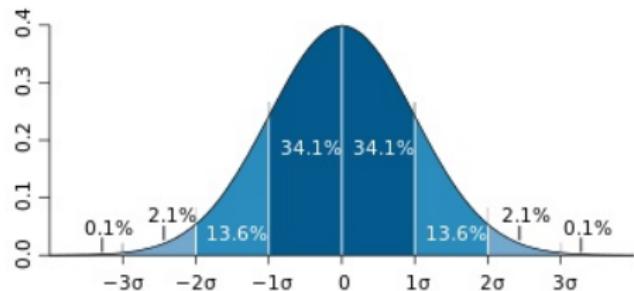
```
1 # Drop the col where all elements are missing values:  
2 df.dropna(axis=1, how='all')  
3  
4 # Drop the col where any of the elements're missing values  
5 df.dropna(axis=1, how='any')  
6  
7 # Keep only the rows which contain 2 missing values max  
8 df.dropna(thresh=2)  
9  
10 # Fill all missing values with the mean of the column  
11 df.fillna(df.mean())  
12  
13 # Fill any missing value in col 'A' with the col median  
14 df['A'].fillna(df['A'].median())
```

Outliers

- Así como hay situaciones donde es importante mantenerlos en otros casos es necesario **eliminarlos**.
- Se recomienda hacer un análisis previo basado en la **naturaleza** de los datos.
- Algunas técnicas comunes son:
 - Removerlos usando desviación estándar (**PYTHON**).
 - Removerlos usando percentiles (**pandas**).



Outliers - Desviación Estándar



Regla en distribución normal: Remover datos en $(mean + 2\sigma)$ y $(mean - 2\sigma)$

Regla 68-95-99 ó 3σ

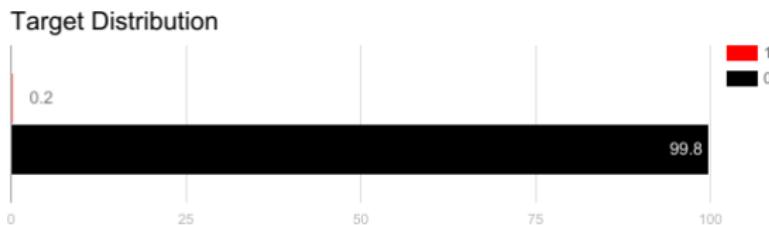
Outliers - Desviación Estándar

PYTHON

```
1 import numpy  
2  
3 arr = [10, 386, 479, 491, 501 ... 411, 399, 363, 19, 543]  
4  
5 elements = numpy.array(arr)  
6  
7 mean = numpy.mean(elements, axis=0)  
8 sd = numpy.std(elements, axis=0)  
9  
10 final_list = [x for x in arr if (x > mean - 2 * sd)]  
11 final_list = [x for x in final_list if (x < mean + 2 * sd)]  
12 print(final_list)
```

Datos no-balanceados

- Ocurre cuando una clase de datos en el conjunto, posee una mayoría importante de la cantidad de datos e.g un conjunto de datos de 2 clases, donde: $CLASE1 = 98\%$ y $CLASE2 = 2\%$.



Datos no-balanceados

- Algunas técnicas comunes son:
 - Usar otras métricas diferentes al porcentaje de exactitud, por ejemplo:
 - Precision/Specificity: cuantas instancias seleccionadas son relevantes.
 - Recall/Sensitivity: cuantas instancias relevantes son seleccionadas.
 - F1 score: media harmónica de “precision” y “recall”.
 - Muestreo de datos:
 - sub-muestreo: Eliminar instancias abundantes (sólo si hay suficientes datos).
 - sobre-muestreo: Generar instancias faltantes (mediante métodos de repetición o generación, solo en caso de que sea posible)
 - Descomponer el conjunto de datos en subconjuntos.
 - Hacer clustering de grupo abundante.

Transformación de datos

- Ocurre cuando **transformamos** un valor z_i en y_i mediante una función $f()$ de forma tal que $y_i = f(z_i)$.
- Se hace con el fin de alinear los datos con alguna suposición estadística, mejorar la interpretación de los datos o bien obtener gráficos de mejor apariencia.
- Técnica muy común: **One Hot Encode**, permite convertir datos categóricos en numéricos (vectores binarios).

Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

Datos Categóricos

Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Vectores Binarios

$$\begin{bmatrix} 0 & 0 & 0 & \boxed{1} & 0 \end{bmatrix} \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ \boxed{10} & 12 & \boxed{19} \\ 11 & 18 & 25 \end{bmatrix} = [10 \ 12 \ 19]$$

One Hot Encode

Algunas Fuentes de DATASETS

- Kaggle Datasets
 - www.kaggle.com
- UCI Machine Learning Repository
 - <https://archive.ics.uci.edu/ml/index.php>
- Google's Datasets
 - <https://toolbox.google.com/datasetsearch>
- Microsoft Datasets
 - <https://msropendata.com/>
- Awesome Public Datasets
 - <https://github.com/awesomedata/awesome-public-datasets>

Questions?



Felipe Meza - fmeza@itcr.ac.cr