Machine Learning

Introducción

Dr. Mauricio Toledo-Acosta

Diplomado Ciencia de Datos con Python 2024

III. Aprendizaje Automático

(6 semanas – 30 hrs.)

1. **Preprocesamiento y métricas**
   a. **Limpieza**
   b. **Normalización**
   c. **Transformación**
   d. **Métricas de rendimiento y funciones de pérdida**
2. **Regresión y Clasificación**
   a. **Regresión lineal**
   b. **Regresión logística**
   c. **Árboles de decisión**
   d. **SVM**
   e. **Bayes Ingenuo**
3. **Agrupamiento y reducción de dimensión**
   a. **K-MEANS**
   b. **Métodos aglomerativos**
   c. **PCA, t-sne**

Table of Contents

¿Qué es la Inteligencia Artificial?

La Inteligencia Artificial (IA) es el conjunto de sistemas o algoritmos, cuyo propósito es crear máquinas que imiten la inteligencia humana para realizar tareas y pueden mejorar conforme la información que recopilan.

> *"... la capacidad de un sistema para interpretar correctamente datos externos, y así aprender y emplear esos conocimientos para lograr tareas y metas concretas a través de la adaptación flexible"*
>
> *A. Kaplan, M. Haenlein, 2019.*

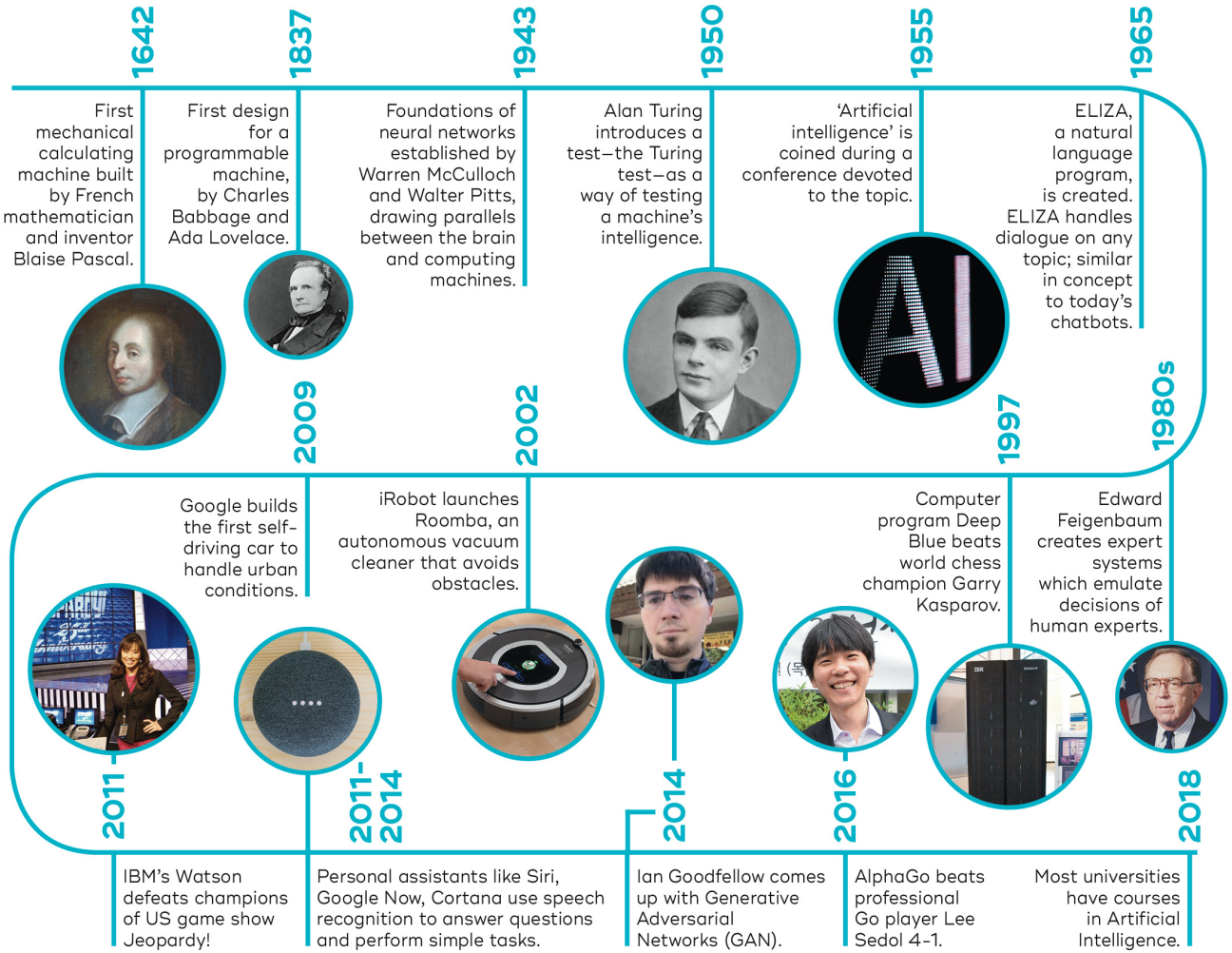
Las definiciones dependen de relaciones con la inteligencia y cognición humana: *aprender*, *interpretar*, *aprender*, ...

Tipos de Inteligencia Artificial

Clasificación de acuerdo a Russell & Norvig, *Artificial Intelligence: A modern approach*, (2002).

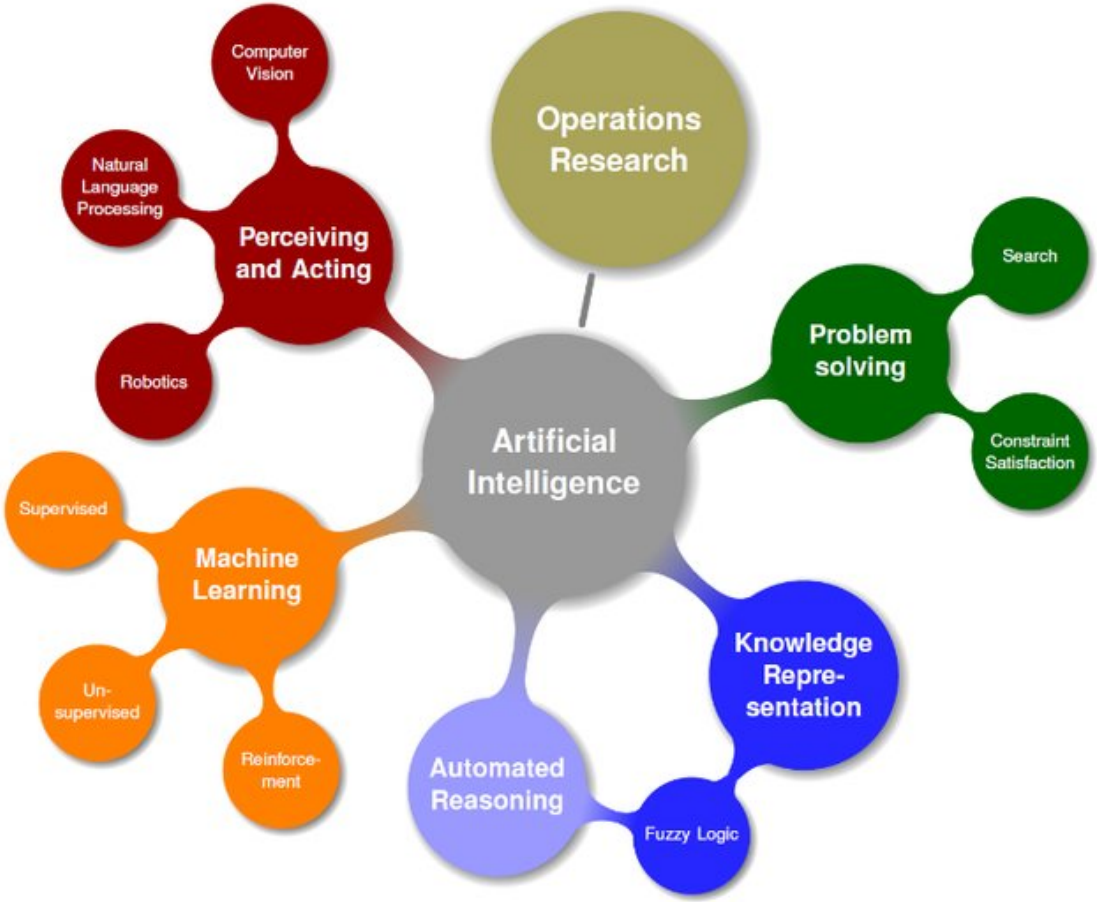
- Los sistemas que piensan como humanos: Estos sistemas tratan de emular el pensamiento humano automatizando actividades como la toma de decisiones, resolución de problemas y aprendizaje.
- Los sistemas que actúan como humanos: Estos sistemas tratan de actuar como humanos; es decir, imitan el comportamiento humano.

Evolución de la IA



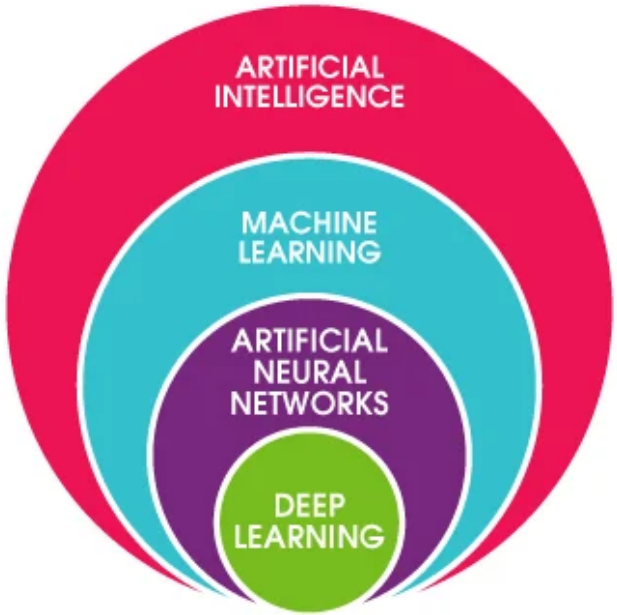
https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence

Algunas áreas de la IA



doi:10.13140/RG.2.2.23097.80485

Algunas áreas de la IA



¿Qué es el Machine Learning?

Machine Learning

El Machine Learning es una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan con el objetivo de resolver ciertas tareas.

¿Qué es el Machine Learning?

Machine Learning

El Machine Learning es una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan con el objetivo de resolver ciertas tareas.

Se dice que un programa de computadora aprende de la Experiencia E con respecto a alguna Tarea T y una medida de desempeño P, si su desempeño sobre T, medido por P, mejora con la experiencia E.

Machine Learning vs Algoritmos tradicionales

Un algoritmo tradicional toma una entrada y una lógica en forma de código y genera una salida para resolver un problema.

Machine Learning vs Algoritmos tradicionales

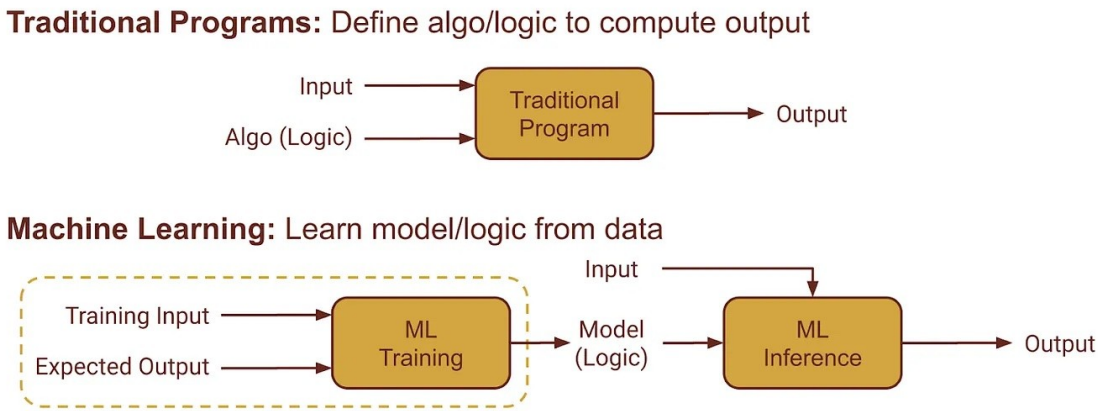
Un algoritmo tradicional toma una entrada y una lógica en forma de código y genera una salida para resolver un problema.

Por el contrario, un algoritmo de aprendizaje automático toma una entrada y una salida y aprende una lógica que puede utilizarse para trabajar con nuevas entradas y obtener una salida. Esta lógica se obtiene a partir de los patrones presentes en los datos.

Machine Learning vs Algoritmos tradicionales

Un algoritmo tradicional toma una entrada y una lógica en forma de código y genera una salida para resolver un problema.

Por el contrario, un algoritmo de aprendizaje automático toma una entrada y una salida y aprende una lógica que puede utilizarse para trabajar con nuevas entradas y obtener una salida. Esta lógica se obtiene a partir de los patrones presentes en los datos.



https://www.linkedin.com/pulse/machine-learning-vs-traditional-software-development-ml4devs-gupta

Fases de un programa de Machine Learning

Los programas de aprendizaje automático tienen dos fases distintas:

1. Entrenamiento: Las entradas y la salida esperada se utilizan para entrenar y probar varios modelos. Se selecciona el modelo más adecuado. *Entrenar* quiere decir determinar los parámetros adecuados del modelo para producir la salida esperada, a partir de las entradas.

2. Inferencia o predicción: El modelo se aplica a nuevos datos de entrada para predecir nuevas salidas, las cuales pueden compararse con las salidas reales.

¿Por qué necesitamos el Machine Learning?

En el enfoque clásico, antes del ML, se usaban algoritmos que procesaban los datos con base en reglas lógicas (`if, else, ...`).

¿Por qué necesitamos el Machine Learning?

En el enfoque clásico, antes del ML, se usaban algoritmos que procesaban los datos con base en reglas lógicas (`if, else, ...`).

- La lógica para tomar las decisiones es específica de acuerdo al dominio y a la tarea. Pequeños cambios en la tarea requieren rediseñar el sistema.
- El diseño de las reglas requiere un entendimiento profundo del dominio por parte de un experto. Estas reglas pueden ser muy complicadas.

¿Por qué necesitamos el Machine Learning?

En el enfoque clásico, antes del ML, se usaban algoritmos que procesaban los datos con base en reglas lógicas (`if, else, ...`).

- La lógica para tomar las decisiones es específica de acuerdo al dominio y a la tarea. Pequeños cambios en la tarea requieren rediseñar el sistema.
- El diseño de las reglas requiere un entendimiento profundo del dominio por parte de un experto. Estas reglas pueden ser muy complicadas.

Ejemplos:

ML	Tradicional
Detectar correo SPAM	Integrar numéricamente una función
Reconocer rostros en una imagen	

Un ejemplo: Detección de SPAM

Enfoque Tradicional

Enfoque Machine Learning

¿Qué tareas puede resolver el Machine Learning?

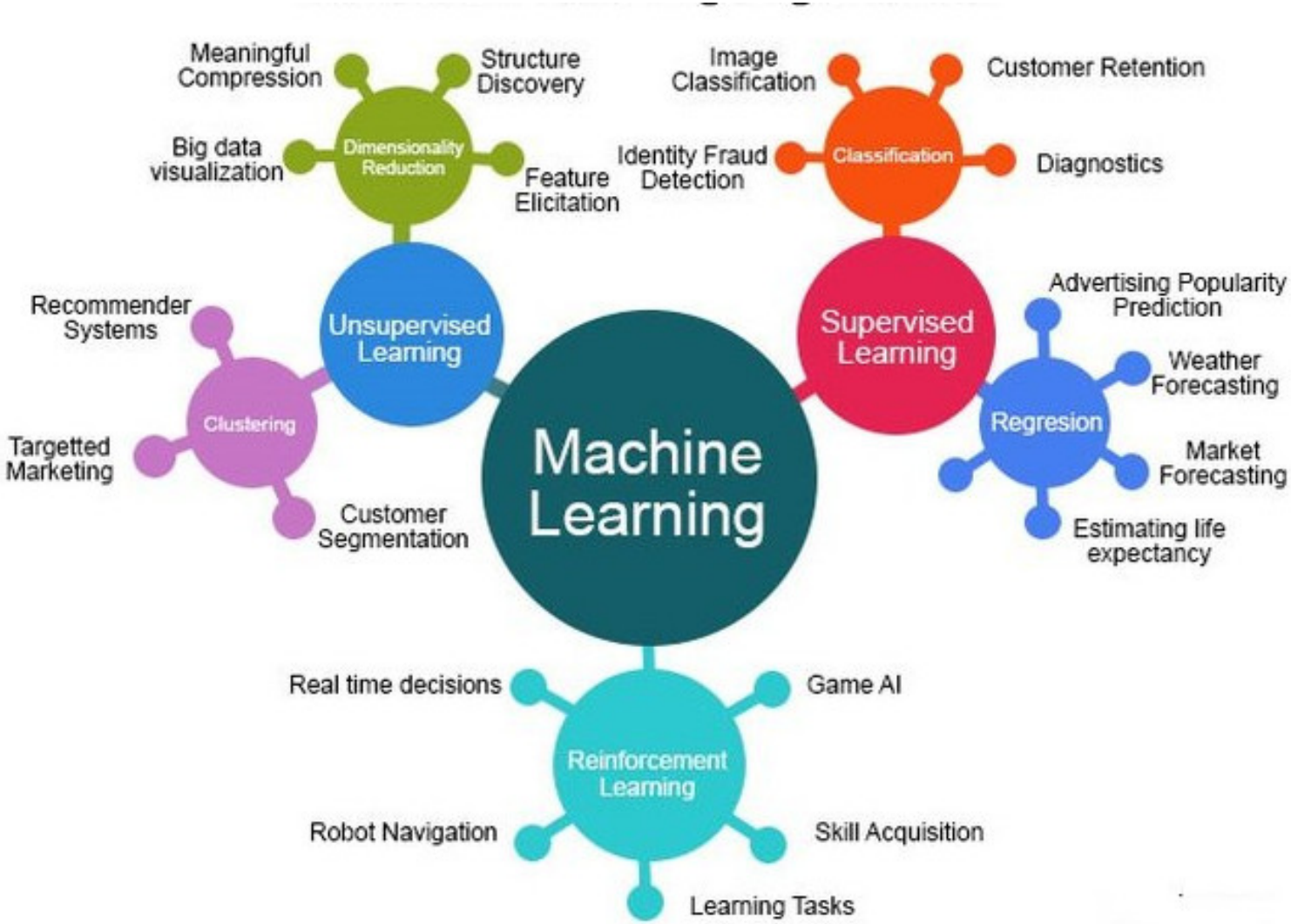
Algunas de las tareas que pueden resolver los métodos de Machine Learning son

1. Regresión.
2. Clasificación.
3. Clustering (segmentación).
4. Traducción automática.
5. Detección de anomalías.
6. Síntesis y muestreo.
7. Generación (texto, imágenes).

Los tres paradigmas del Machine Learning

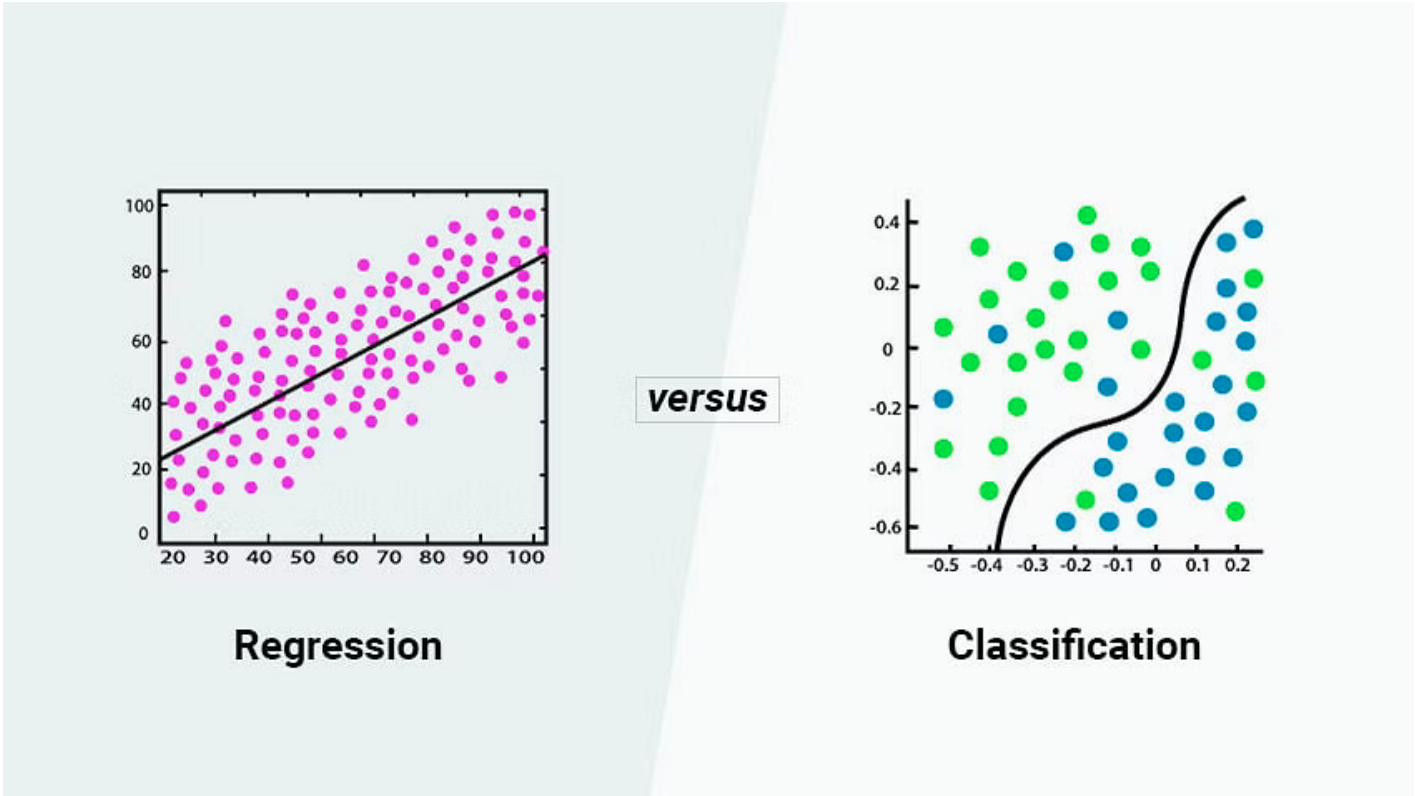
1. Aprendizaje supervisado (Supervised Learning). El modelo aprende de ejemplos etiquetados para hacer predicciones de nuevos datos.
2. Aprendizaje no supervisado (Unsupervised Learning). El modelo encuentra patrones o estructuras intrínsecas en los datos sin etiquetar.
3. Aprendizaje por refuerzo (Reinforcement Learning). El modelo aprende las acciones a tomar en un entorno para maximizar una noción de recompensa acumulativa.

Los tres paradigmas del Machine Learning



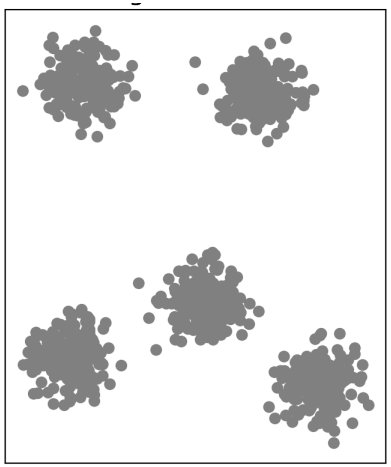
Aprendizaje Supervisado

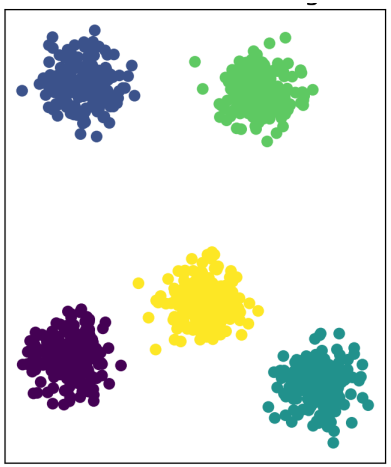
1. Regresión.
2. Clasificación.



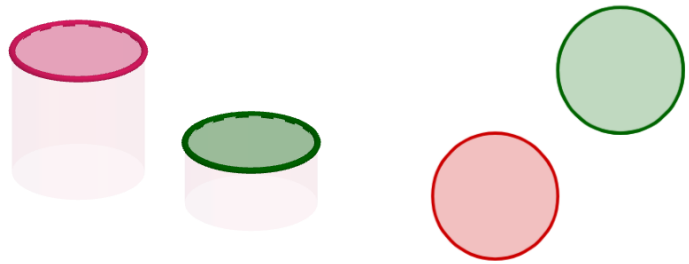
Aprendizaje No Supervisado

Clustering





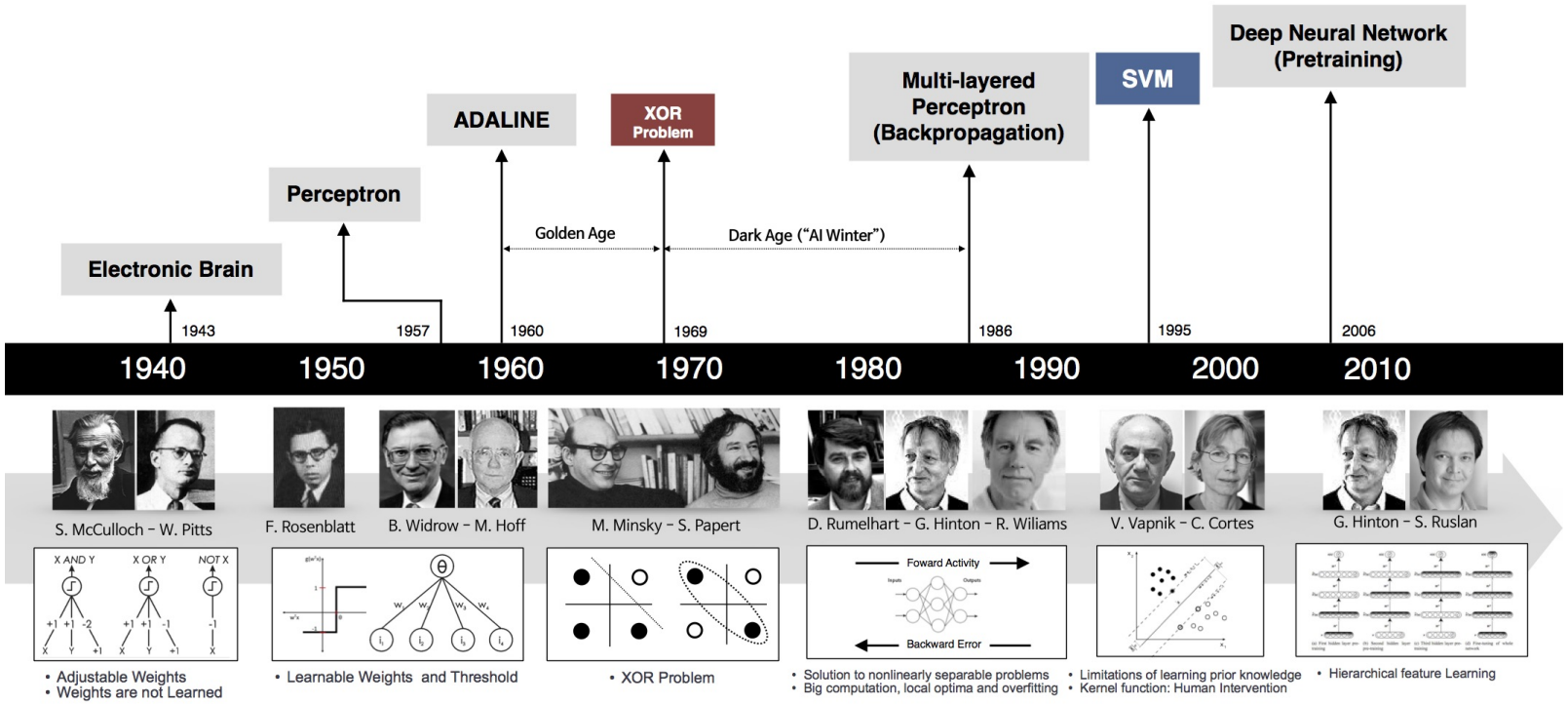
Reducción de dimensionalidad



Aprendizaje por Refuerzo

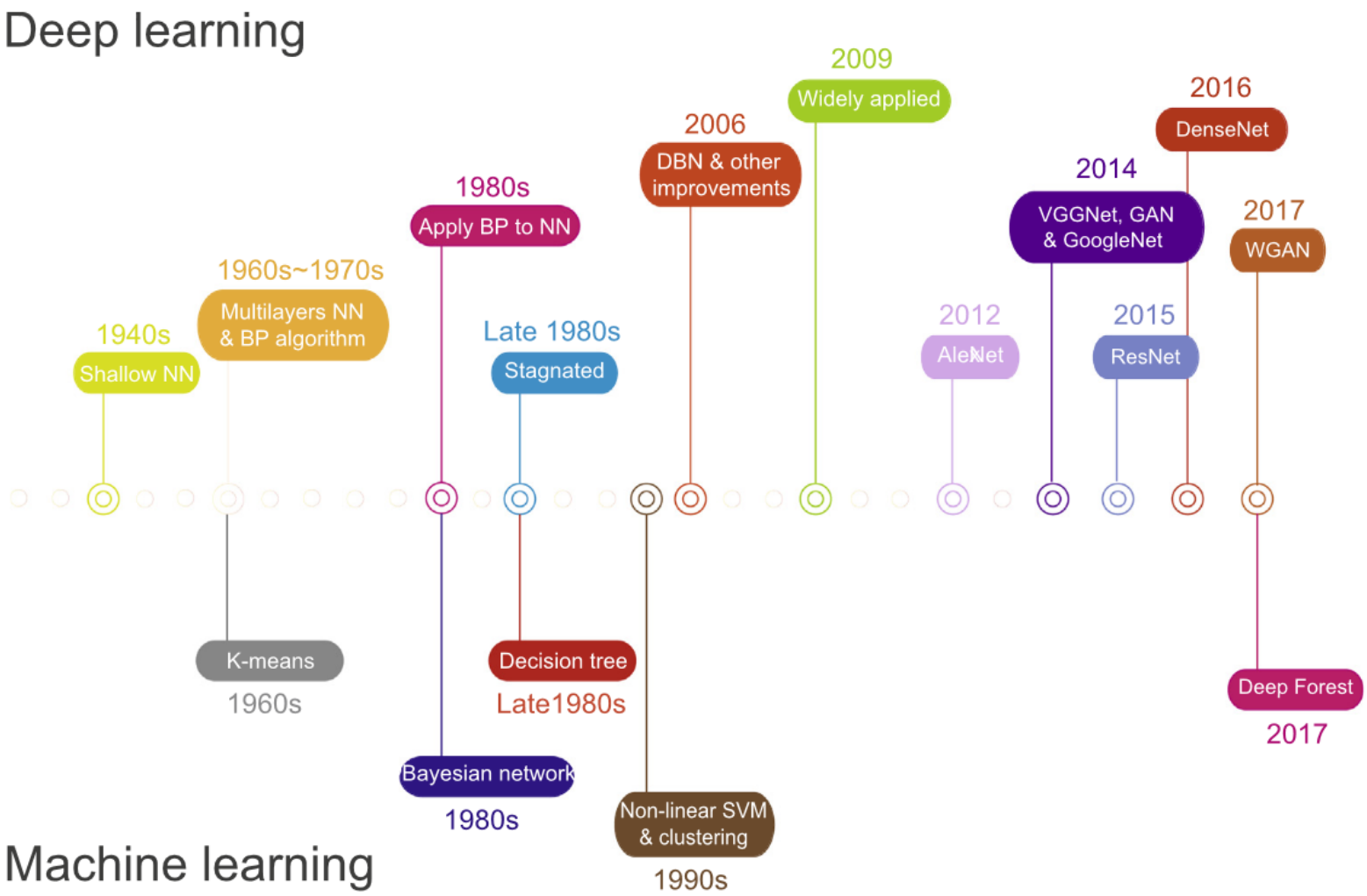
1. Navegación en vehículos autónomos.
2. Texto predictivo.
3. Sistemas de recomendación.
4. Videojuegos.
5. Conservación y eficiencia energética.

Timeline de los algoritmos de Machine Learning



http://beamlab.org/deeplearning/2017/

Timeline de los algoritmos de Machine Learning



https://doi.org/10.1016/j.gpb.2017.07.003

Referencias

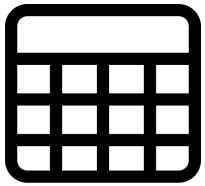
- Müller, A. C., & Guido, S., 2016. *Introduction to Machine Learning with Python: a Guide for Data Scientists*. O'Reilly Media, Inc..
- Flach, P. A., 2012. *Machine Learning : the Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press.

Table of Contents

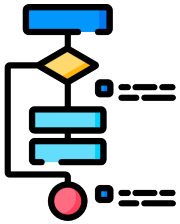
Componentes del Machine Learning

Datos

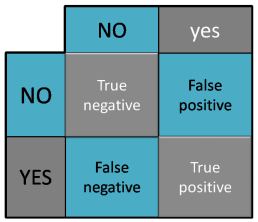
Variables (features)



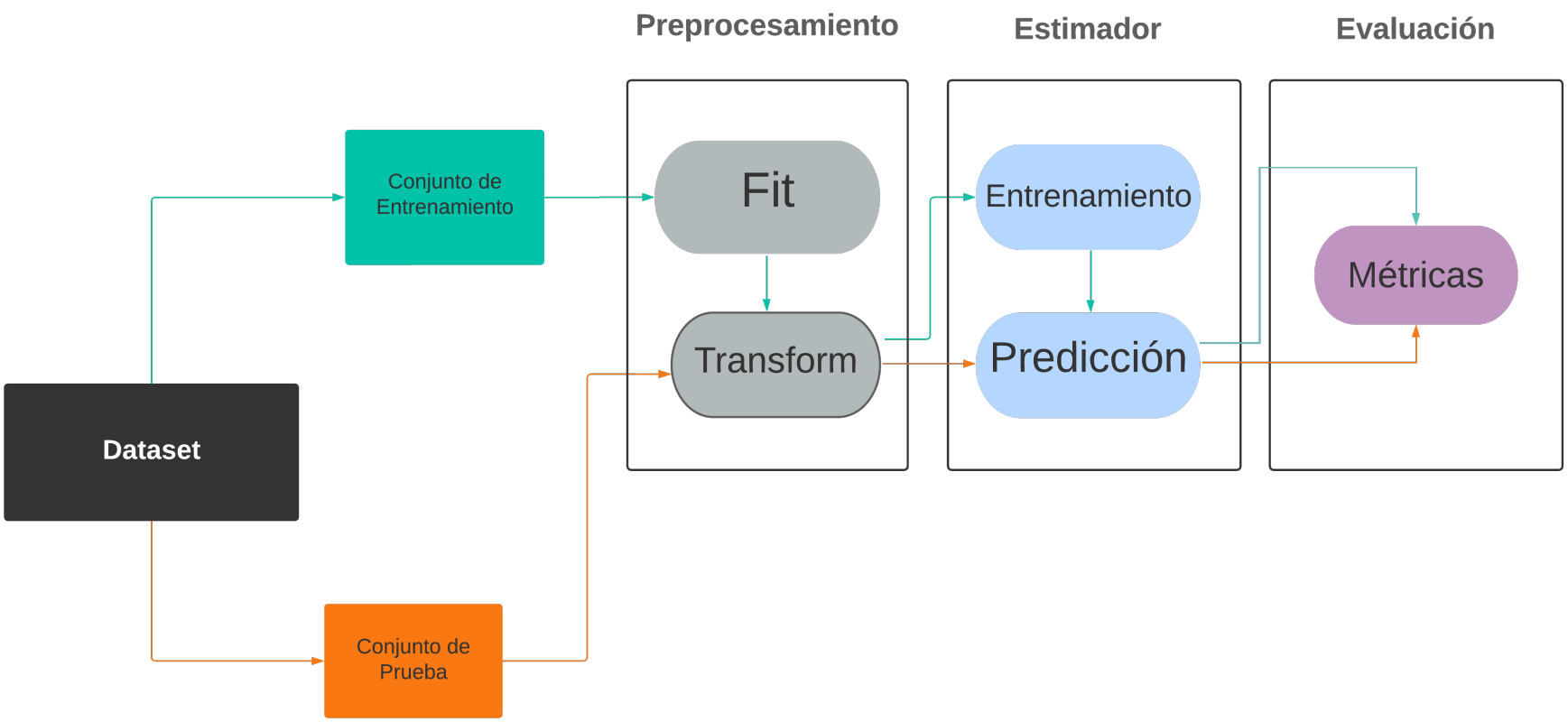
Algoritmos



Métricas



Workflow del Machine Learning



Datos

Los datos pueden tener muchas formas diferentes:

- Tablas estructuradas.
- Imágenes.
- Texto.
- Archivos de audio.
- Archivos de video.

A un conjunto de datos, se le llama dataset. Kaggle tiene una colección grande, al igual que Scikit-Learn.

$$\text{Dataset} \neq \text{Base de datos}$$

Algunos datasets famosos

- MNIST
- Iris Flowers Dataset.
- Boston House Price Dataset.
- Wine Quality Dataset.
- Pima Indians Diabetes Dataset.
- 20newsgroups.

Features

Los datasets suelen ser tablas donde cada fila representa una entidad y cada columna una característica de esa entidad. Es decir, cada entidad esta representada por un conjunto de caracteristicas.

Hay varios tipos de caracteristicas (variables):

- Numéricas
  - Continuas: 9.32, $\sqrt{2}$, ...
  - Discretas: 1,2,3,...
- Categóricas
  - Ordinales: escalas numéricas.
  - Nominales: género, medio de transporte, tipos.

Preprocesamiento

En cualquier proceso de Machine Learning, el preprocesamiento (PP) es el paso en el que los datos se transforman o codifican para llevarlos a un estado tal que ahora la máquina pueda analizarlos de *mejor forma*.

Preprocesamiento

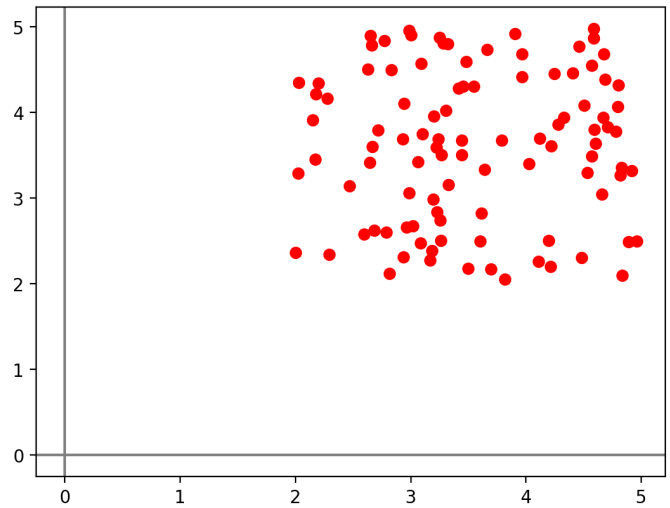
En cualquier proceso de Machine Learning, el preprocesamiento (PP) es el paso en el que los datos se transforman o codifican para llevarlos a un estado tal que ahora la máquina pueda analizarlos de *mejor forma*.

Estos son algunos de los tipos de problemas básicos así como la familia de técnicas de PP a la que pertenecen:

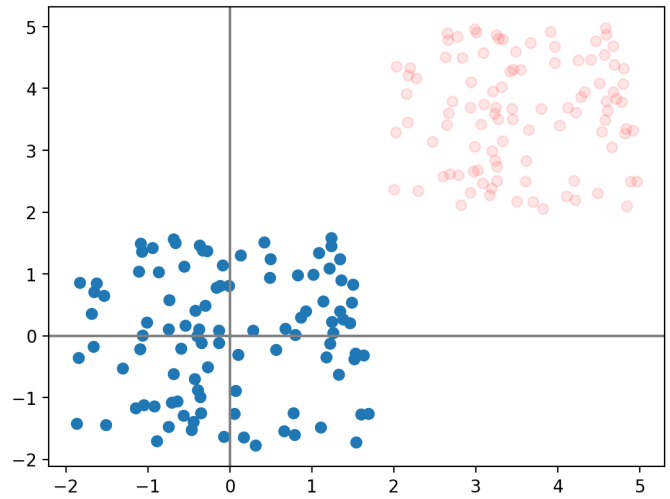
- ¿Cómo limpio los datos? Limpieza de datos.
- ¿Cómo unifico y escalo los datos? Normalización de datos.
- ¿Cómo proporciono datos precisos? Transformación de datos.
- ¿Cómo manejo los datos faltantes? Imputación de datos perdidos.
- ¿Cómo incorporo y ajusto datos? Integración de datos.
- ¿Cómo detecto y manejo el ruido? Análisis del ruido.

Preprocesamiento

Datos originales



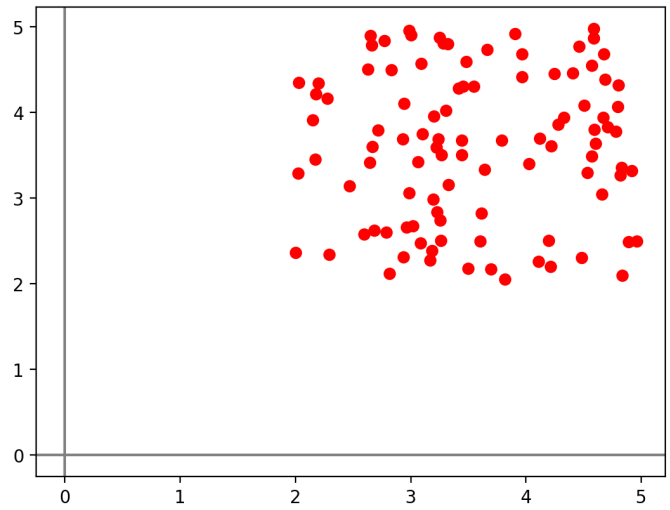
Preprocesamiento



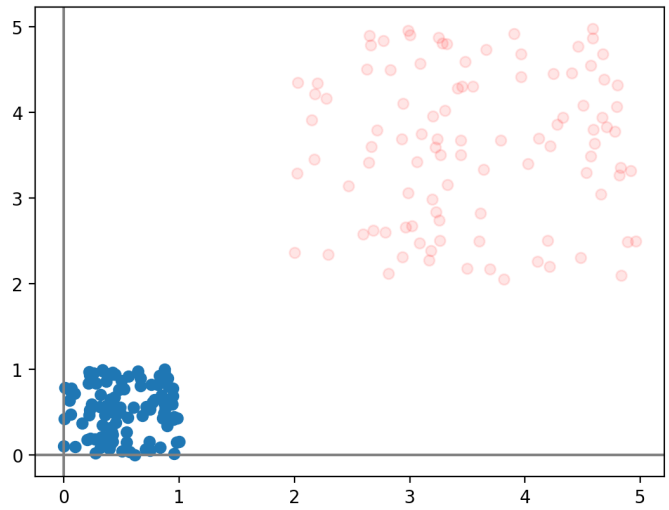
`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

Preprocesamiento

Datos originales



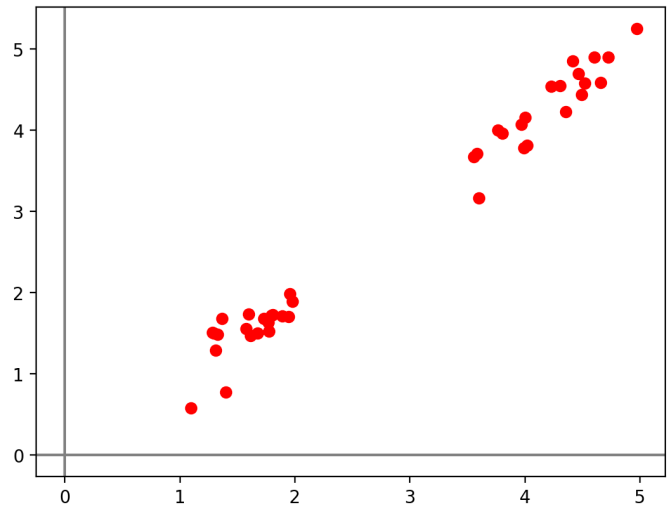
Preprocesamiento



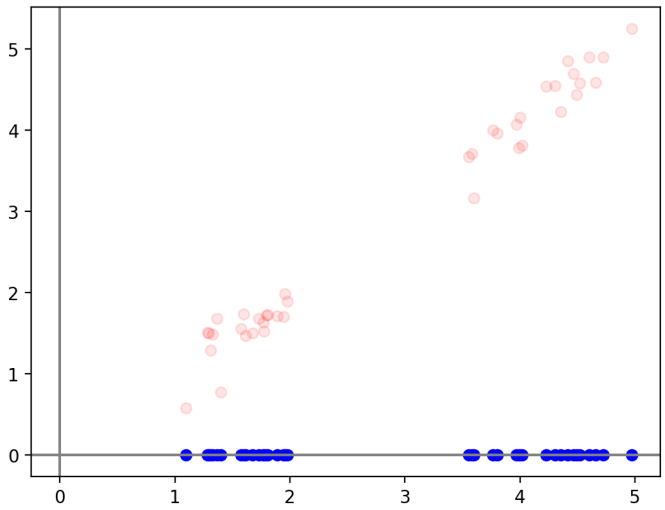
`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

Preprocesamiento

Datos originales

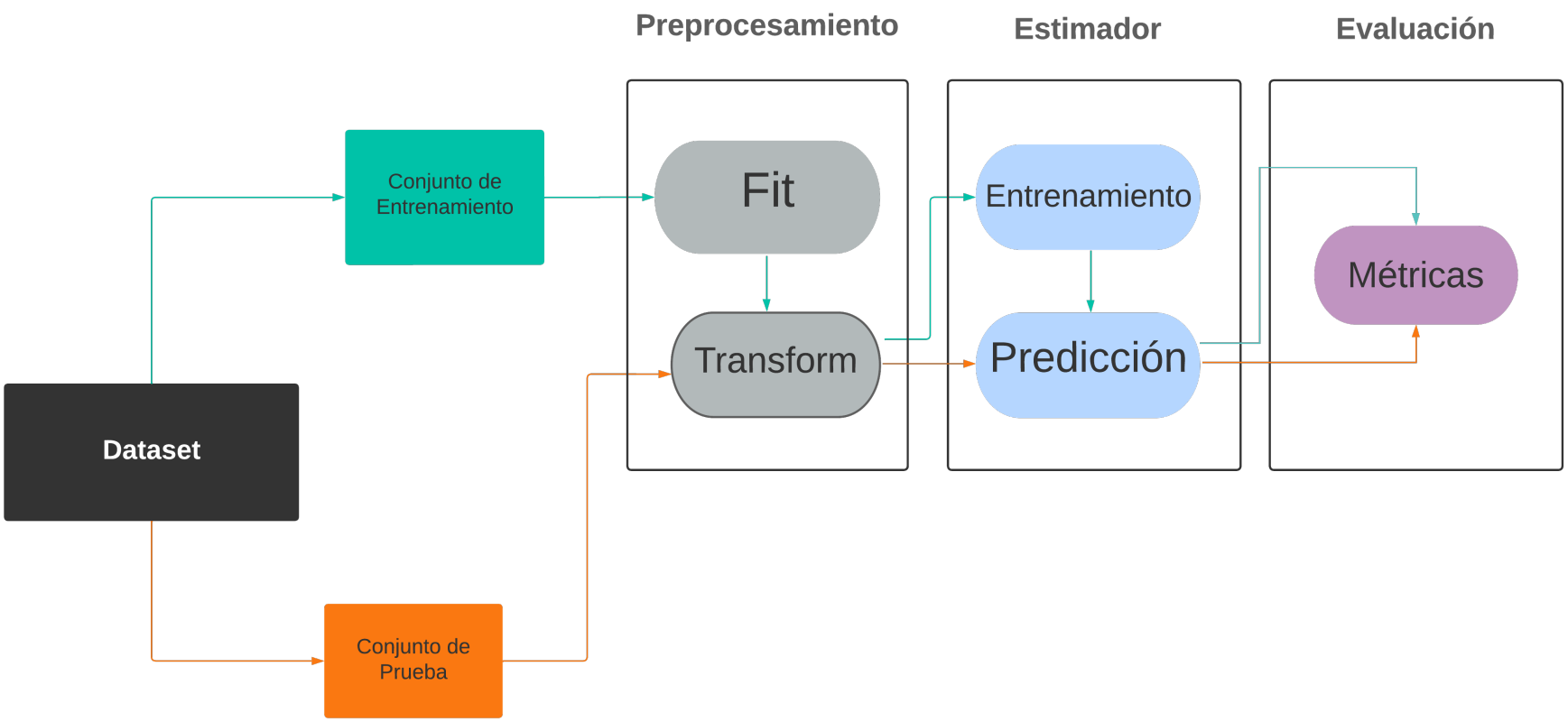


Preprocesamiento



`https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html`

Workflow del Machine Learning: Algoritmos



Algoritmos

Un algoritmo de Machine Learning es la técnica que permite a una computadora aprender a partir de datos y tomar decisiones o hacer predicciones basadas en esa información.

Algoritmos

Un algoritmo de Machine Learning es la técnica que permite a una computadora aprender a partir de datos y tomar decisiones o hacer predicciones basadas en esa información.

Existen varios tipos de algoritmos de Machine learning, dependiendo del tipo de tarea que buscar modelar:

- Aprendizaje supervisado.
- Aprendizaje no supervisado.
- Aprendizaje por refuerzo.

Algoritmos


100+ MACHINE LEARNING ALGORITHMS
By @TheInsaneApp
CLUSTERING
- k Means
- k Medians
- BIRCH
- Fuzzy C-Means
- Fuzzy K-Modes
- Mini Batch K-Means
- DBSCAN
- Fuzzy clustering
- Mean-shift
- Optics algorithm
- Expectation Maximization
- Hierarchical Clustering
- Minimum spanning tree
DIMENSIONALITY REDUCTION
- PCA Algorithm
- PCR Algorithm
- PLSR Algorithm
- Sammon Mapping
- MDS Algorithm
- Projection Pursuit
- LDA Algorithm
- ICA Algorithm
- NMF Algorithm
- RDA Algorithm
- MDA Algorithm
- PLSDA Algorithm
- QDA Algorithm
- CCA Algorithm
- Diffusion Map
REGRESSION
- Ordinary Least Squares Regression
- Linear Regression
- Logistic Regression
- Stepwise Regression
- MARS Algorithm
- Locally Estimated Scatterplot Smoothing
DECISION TREE
- Conditional Decision Trees
- Decision Stump
- C4.5 and C5.0
- CHAID
- CART
- ID3
- M5
BAYESIAN
- AODE
- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bayesian Belief Network
- Bayesian Network
REINFORCEMENT LEARNING
- Q Learning
- SARSA Algorithm
- Deep Q-network
- Learning Automata
- DDPG Algorithm
- NAF Algorithm
- A3C Algorithm
- TRPO Algorithm
- PPO Algorithm
- Constructing Skill Trees
REGULARIZATION
- Ridge Regression
- Elastic Net
- LASSO
- LARS
EMSEMBLE
- GBDT
- GBRT
- Boosting
- Bagging
- AdaBoost
- Random Forest
- Blending Algorithm
- Stacked Generalization
- Gradient Boosting Machines
ANNs
- Perceptron
- MLP Algorithm
- Back-Propagation
- Stochastic Gradient Descent
- Hopfield Network
- RBFN Algorithm
OTHER ML ALGORITHMS (Part 1)
- ALOPEX
- CN2 algorithm
- FastICA
- Feature selection Algorithms
- Linde Buzo Gray Algorithm
- Forward Backward algorithm
- Algorithm Accuracy Evaluation
- Performance Measures
- Optimization Algorithm
- Dynamic Time Warping
- Local Outlier Factor
- Logic Learning Machine
- Markov Chain Monte Carlo
- T Distributed Stochastic neighbor Embedding
- Wake Sleep Algorithm
- Prefrontal Cortex Basal Ganglia Working Memory
RULE SYSTEM
- Cubist
- OneR
- ZeroR
- Ripper
OTHER ML ALGORITHMS (Part 2)
- LogitBoost
- Sparse PCA
- Structured kNN
- WMA Algorithm
- GeneRec
- Leabra
- RProp
DEEP LEARNING
- CNN
- RNNs
- LSTMs
- Stacked Auto Encoders
- Deep Boltzmann Machine
- Deep Belief Networks
INSTANCE BASED
- Learning Vector Quantization
- K Nearest Neighbor
- Self-Organizing Map
- Locally Weighted Learning
- Support Vector Machines
ARL
- Apriori Algorithm
- Eclat Algorithm
- FP Growth Algorithm

Cross Validation

Validación Cruzada

La validación cruzada es una técnica de validación de modelos para evaluar cómo se generalizarán los resultados de un análisis estadístico a un conjunto de datos independiente. La validación cruzada es un método de remuestreo que utiliza diferentes partes de los datos para probar y entrenar un modelo en diferentes iteraciones.

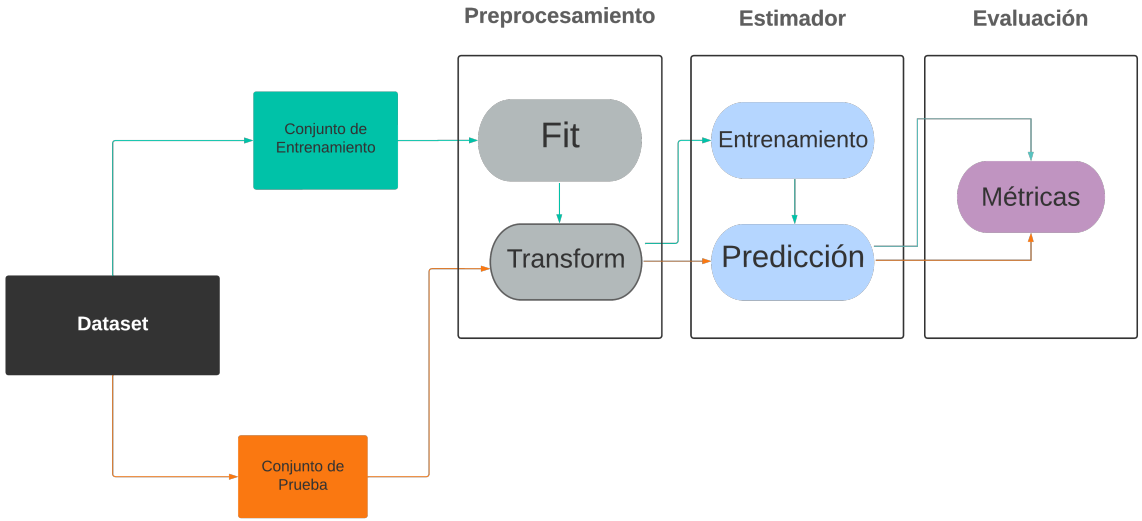
Es necesario tener una validación de la estabilidad de cualquier modelo de Machine Learning. Es decir, ¿qué tan bien podemos esperar que sea su rendimiento en datos que no ha visto?

Tecnicas de validación

- **Validación**: Evaluación del desempeño del modelo en los datos de entrenamiento.

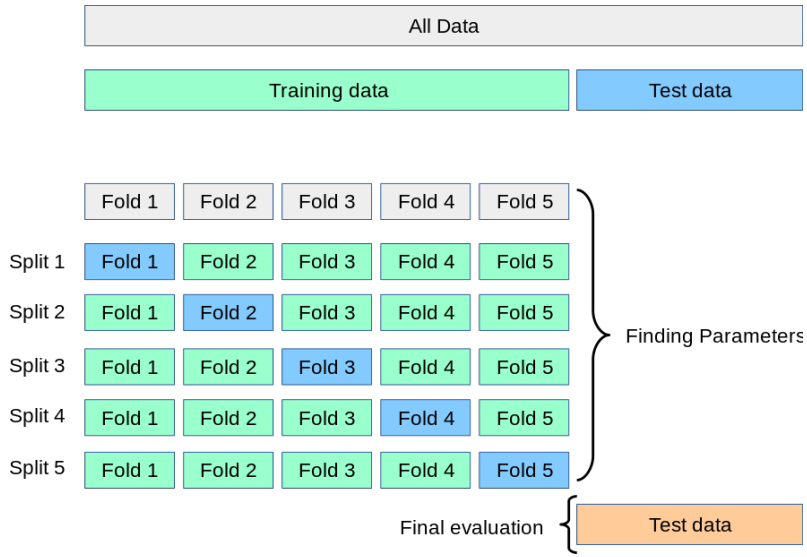
Tecnicas de validación

- **Validación**: Evaluación del desempeño del modelo en los datos de entrenamiento.
- **Conjunto de prueba**: Reservar una parte del conjunto de datos para ser usada como conjunto de prueba.



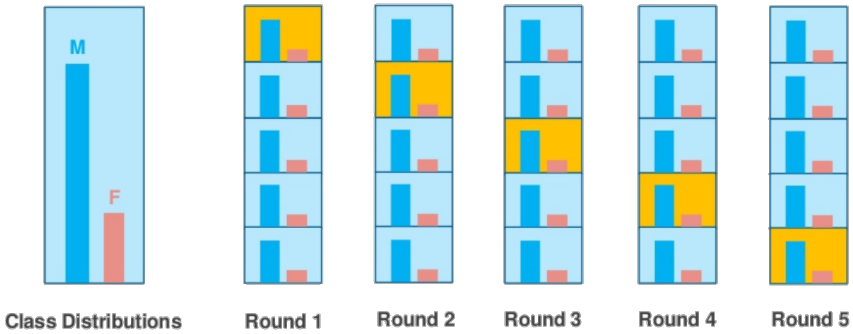
Tecnicas de validación

- **K-Fold Cross Validation**: Los datos se dividen en k subconjuntos, una de las partes se usa como conjunto de prueba y las demás como entrenamiento. Se repite este método k veces, de forma que cada vez, uno de los k subconjuntos se utiliza como conjunto de prueba y los otros $k-1$ subconjuntos, como conjunto de entrenamiento. La estimación del error se promedia sobre las k pruebas para obtener la eficacia total de nuestro modelo.

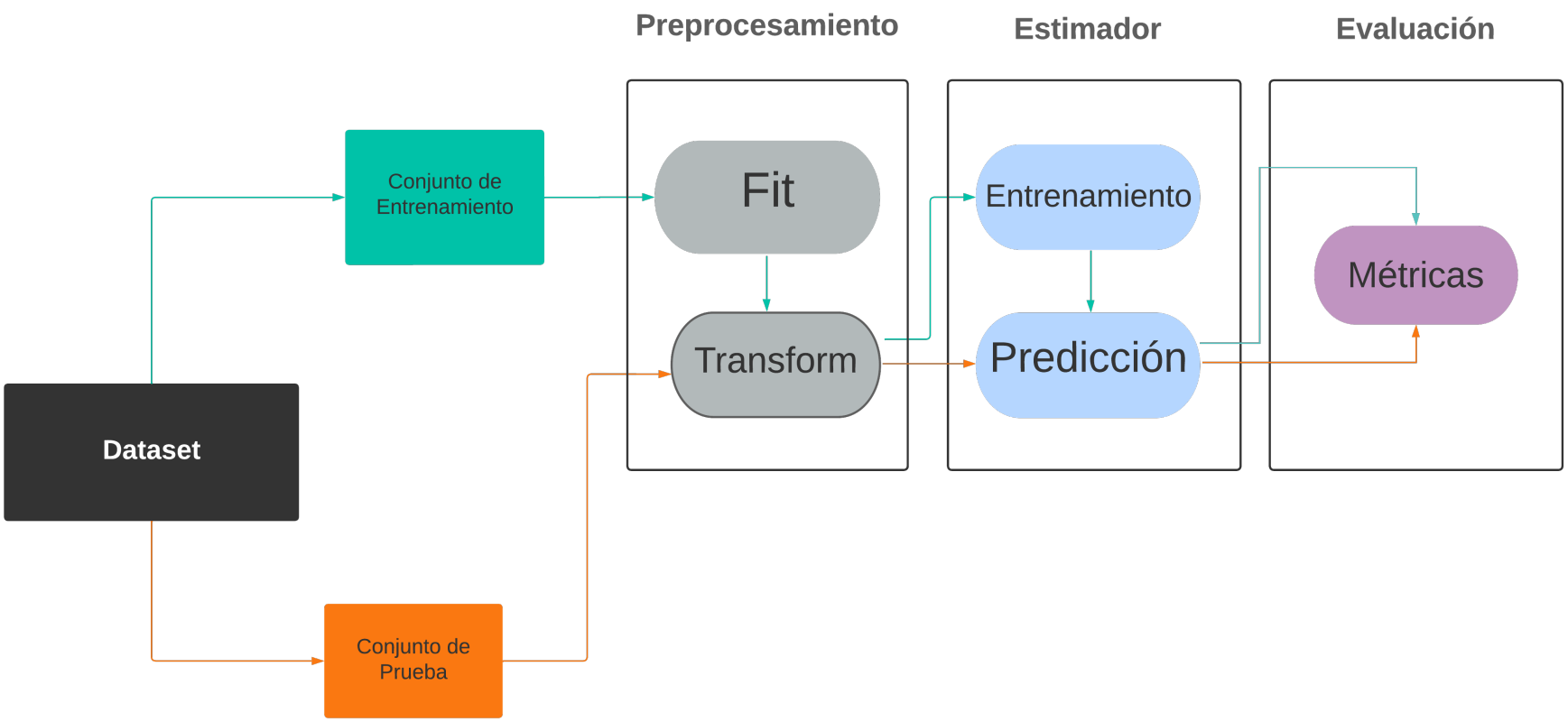


Tecnicas de validación

- **Stratified K-Fold Cross Validation**: Variación de la validación cruzada K-fold normal, en lugar de que las divisiones sean completamente aleatorias, la proporción entre las clases objetivo es la misma en cada uno de los k subconjuntos que en el conjunto de datos completo.



Workflow del Machine Learning: Métricas de rendimiento



Métricas de desempeño: Ejemplo de clasificación

Hello Friends! We hope you had a pleasant week. Last weeks trivia questions was: What do these 3 films have in common: One Crazy Summer, Whispers in the Dark, Moby Dick? Answer: Nantucket Island	?
IMPORTANT INFORMATION: The new domain names are finally available to the general public at discount prices. Now you can register one of the exciting new .BIZ or .INFO domain names, as well as the original .COM and .NET names for just $14.95. These brand new domain extensions were recently approved by ICANN and have the same rights as the original .COM and .NET domain names. The biggest benefit is of-course that the .BIZ and .INFO domain names are currently more available. i.e. it will be much easier to register an attractive and easy-to-remember domain name for the same price. Visit: http://www.affordable-domains.com today for more info.	?
If you have an internal zip drive (not sure about external) and you bios supports using a zip as floppy drive, you could use a bootable zip disk with all the relevant dos utils.	?

Métricas de desempeño: Ejemplo de clasificación

Hello Friends! We hope you had a pleasant week. Last weeks trivia questions was: What do these 3 films have in common: One Crazy Summer, Whispers in the Dark, Moby Dick? Answer: Nantucket Island	No Spam
IMPORTANT INFORMATION: The new domain names are finally available to the general public at discount prices. Now you can register one of the exciting new .BIZ or .INFO domain names, as well as the original .COM and .NET names for just $14.95. These brand new domain extensions were recently approved by ICANN and have the same rights as the original .COM and .NET domain names. The biggest benefit is of-course that the .BIZ and .INFO domain names are currently more available. i.e. it will be much easier to register an attractive and easy-to-remember domain name for the same price. Visit: http://www.affordable-domains.com today for more info.	Spam
If you have an internal zip drive (not sure about external) and you bios supports using a zip as floppy drive, you could use a bootable zip disk with all the relevant dos utils.	No Spam

Métricas de desempeño: Ejemplo de clasificación

Métricas de desempeño: Ejemplo de regresión

Altura	Peso
169.948447	**?**
173.865754	**?**
174.661475	**?**
170.597762	**?**

Métricas de desempeño: Ejemplo de regresión

Altura	Peso
169.948447	82.552060
173.865754	75.674023
174.661475	84.338528
170.597762	87.721204

Métricas de desempeño

Las métricas de desempeño dan cuenta del desempeño del modelo entrenado. Estas funciones varían de acuerdo al tipo de tarea, suelen ser funciones *fácilmente* interpretables (porcentajes, conteos, diferencias, etc.).

- Regresión: MSE, MAE.
- Clasificación: Accuracy, precision, recall, F1-score, ROC-AUC.
- Clustering: AMI, MI, silhoutte score.
- Tareas de NLP: Preplexity, entropy, coherence.
- ...

Funciones de costo

Una función de pérdida, o función de costo, es una función que asigna un evento o los valores de una o más variables a un número real que representa intuitivamente algún *costo* asociado al evento. Un problema de optimización trata de minimizar una función de pérdida.

Un algoritmo de Machine Learning busca minimizar o maximizar esta función cambiando sus parámetros internos. Frecuentemente se usa el descenso de gradiente para este fin, por lo tanto, tipicamente se requiere de una función de costo diferenciable o convexa.

- Regresión: MSE, RMSE, MAE.
- Clasificación: 0-1, binaria asímetrica, entropia cruzada, Hinge loss.

Diferencia entre función de costo y métrica de desempeño

Típicamente son funciones diferentes, bajo ciertas condiciones se puede usar la misma.

- Usando la función de costo como métrica de desempeño: puede ser confusa de interpretar.
- Usando la métrica de desempeño como función de costo: puede no ser posible si no es diferenciable o convexa.

Resumiendo

Un problema de Machine Learning consiste en los siguientes pasos:

- **Recopilación de datos:** Los datos deben ser suficientes y representativos del problema que se busca resolver.
- **Preprocesamiento:** Limpiar los datos para eliminar ruido, valores faltantes, valores atípicos, y los prepara para su uso en el modelo.
- **Selección del algoritmo**.
- **Entrenamiento del modelo:** Utiliza el conjunto de datos de entrenamiento para entrenar el modelo elegido. La mejora se rige usando la función de costo.
- **Evaluación del modelo:** Evalúa el modelo utilizando el conjunto de datos de prueba. Esto se hace con la métrica de rendimiento.
- **Implementación, Monitoreo y Mantenimiento.**

Table of Contents

¿Por qué Python?

Ventajas

- Rápido de aprender.
- El código es claro y fácil de leer.
- Desarrollo rápido de modelos.
- Lenguaje orientado a objetos.
- Muchas librerias: Numpy, Pandas, Scipy, Matplotlib.

Machine Learning

Deep Learning

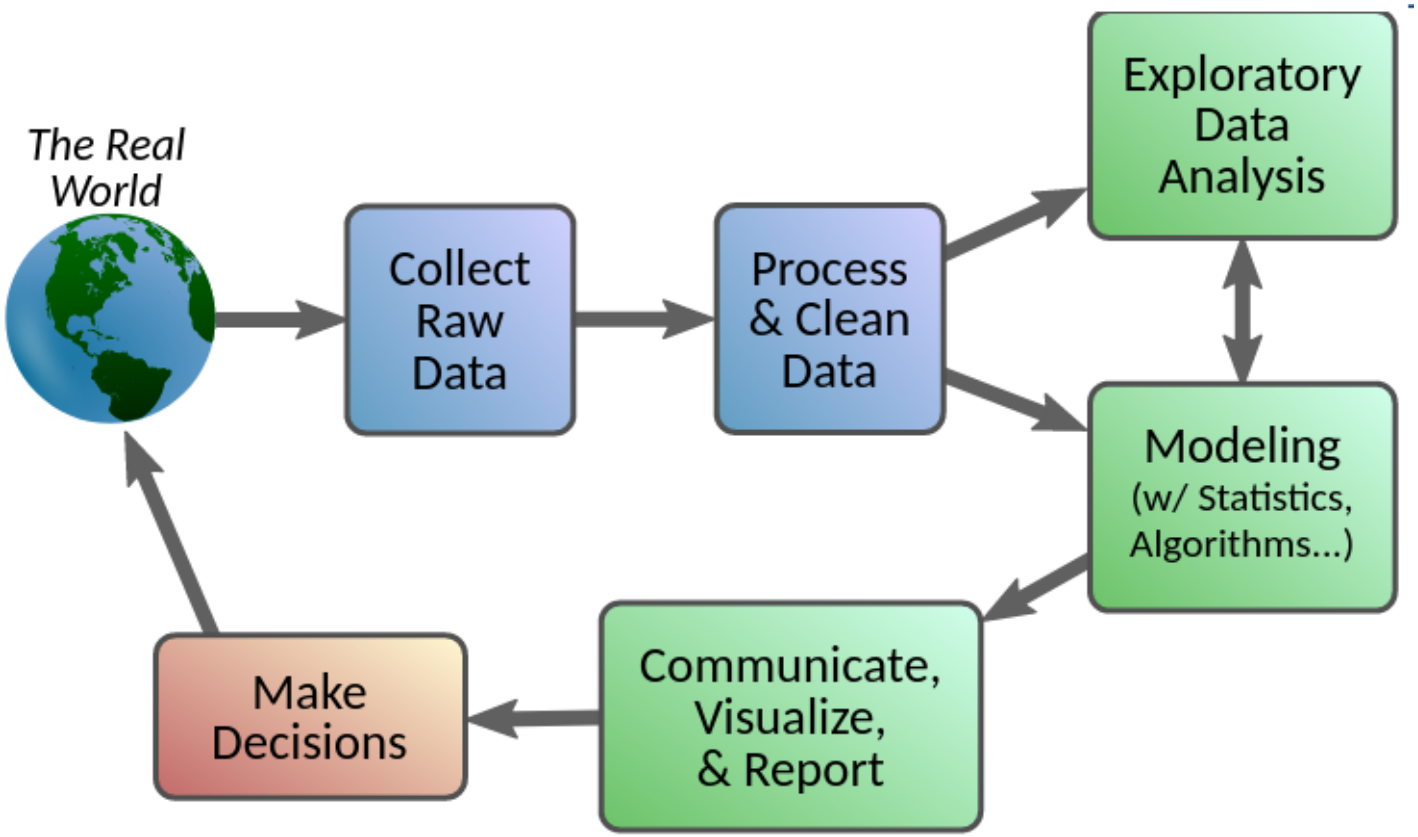
Ciencia de Datos

Ciencia de datos

La ciencia de datos es una disciplina que analiza grandes cantidades de datos para extraer información y patrones que sea útiles en la creación de estrategias que permitan aumentar la eficiencia, reconocer nuevas oportunidades de mercado y aumentar la ventaja competitiva de una organización.

La ciencia de datos emplea las disciplinas de las matemáticas, estadística y las ciencias de la computación. Además, incorpora técnicas del Machine Learning, la mineria de datos y la visualización, entre otras.

El proceso de la ciencia de datos



https://snakebear.science/01-Introduction/WhatIsDS.html

¡Vamos a comenzar!

Welcome to the world of Machine Learning