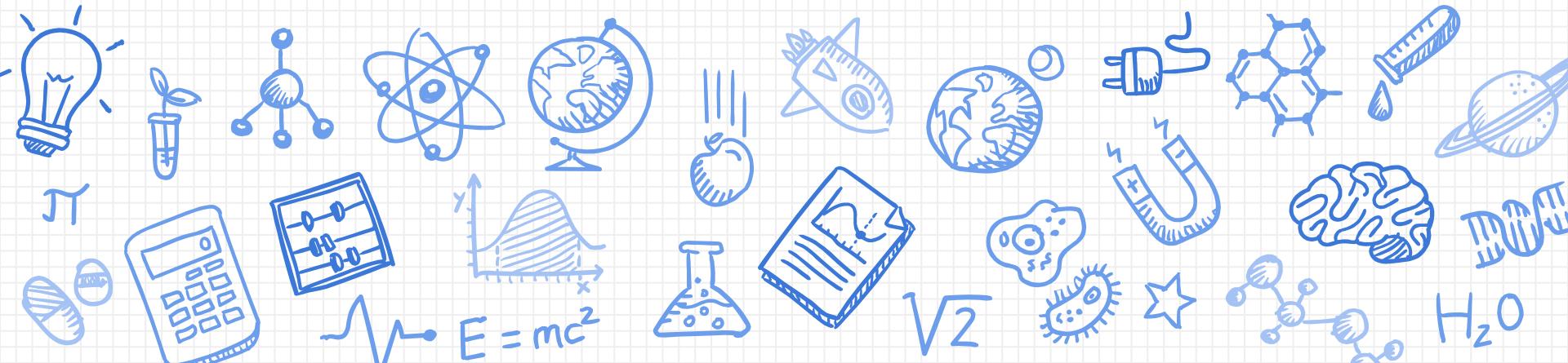


Data Science en la Industria



Soy Cristhian Boujon

- Ingeniero en Sistemas de Información.
- Sr. Software Engineer en ecommerce.
- Profesor.



Desarrollo web hace 10 años atrás...

HTTP

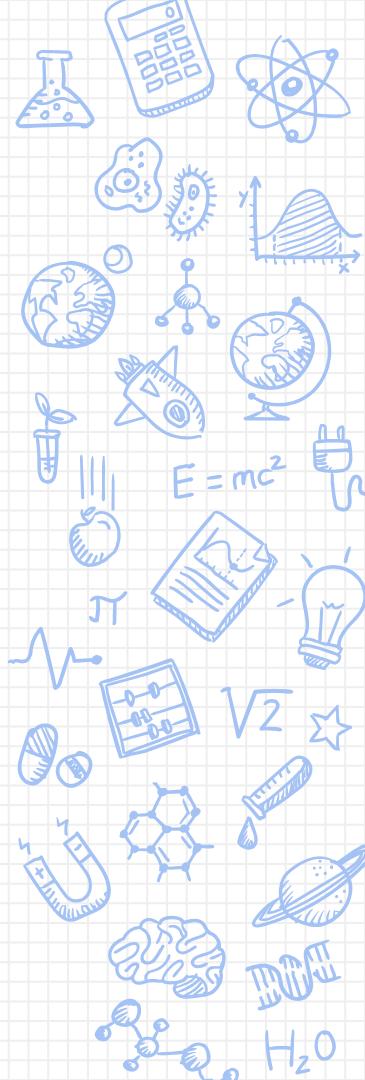
HTML

CSS

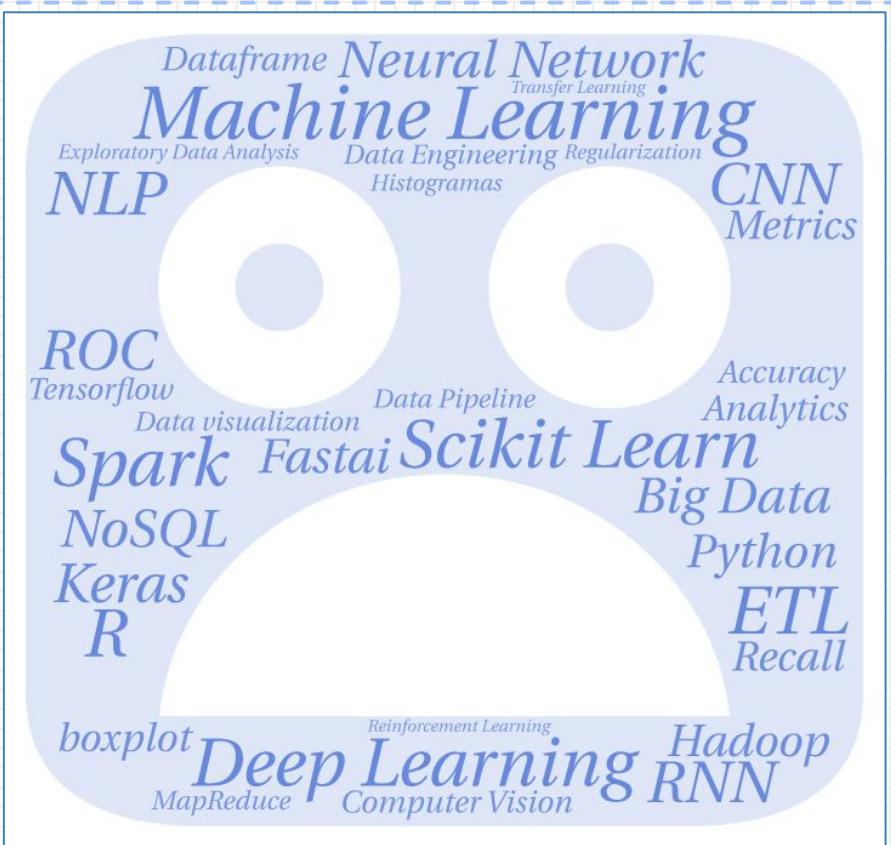
Javascript

PHP

MySQL

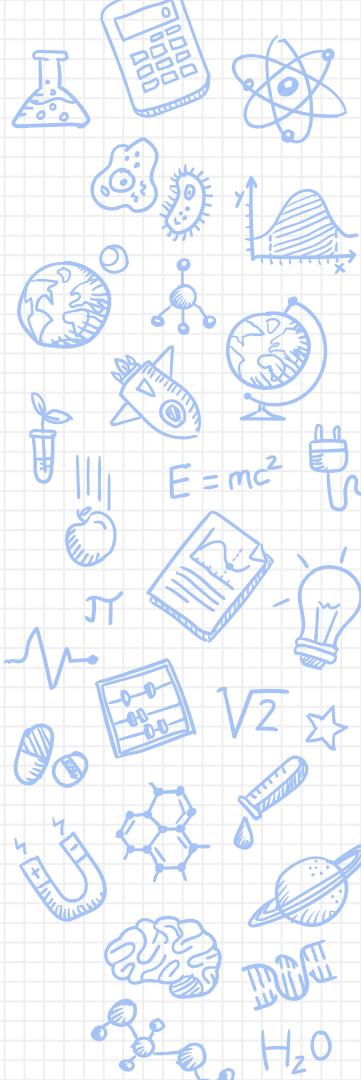


Data Science



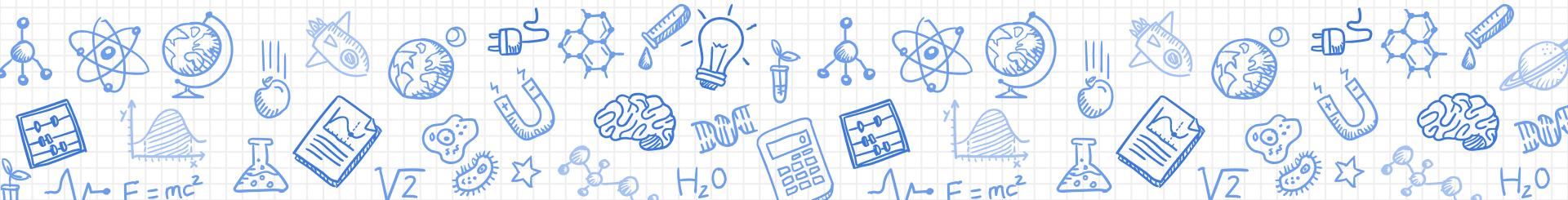
Data Science en la Industria

- Análisis de datos / Data Analysis
 - Modelado / Modeling
 - Ingeniería de datos / Data Engineering



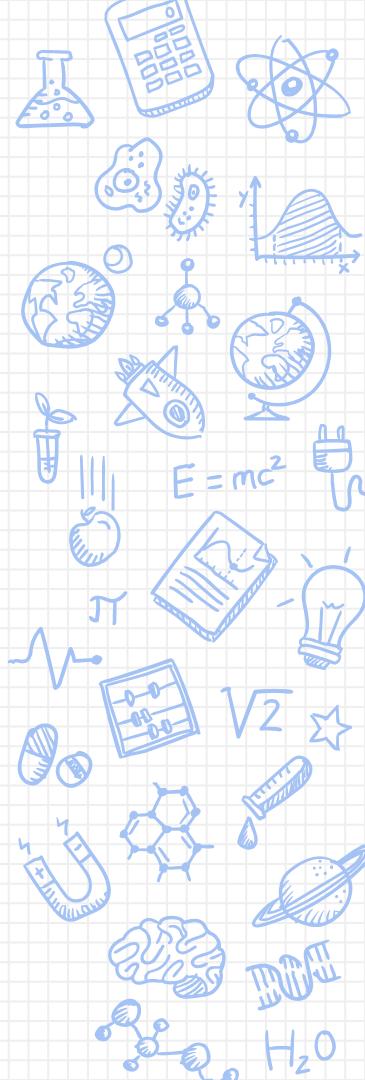
Análisis de Datos

Se examinan los datos en bruto y así sacar conclusiones sobre la información.



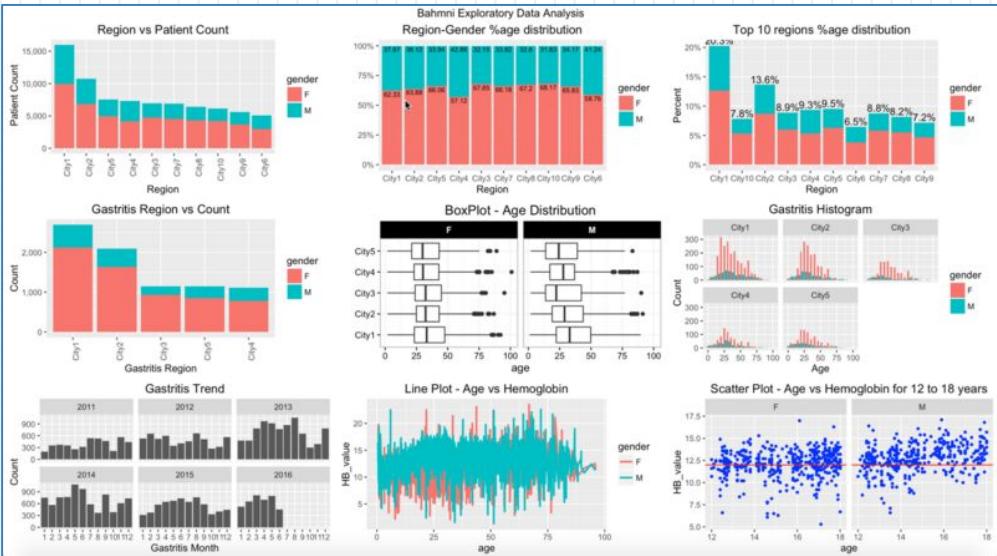
Análisis de Datos – Conocimiento de Dominio

- Conocimientos de Negocio.
 - Necesidad del negocio -> Preguntas
 - Planteo de Hipótesis

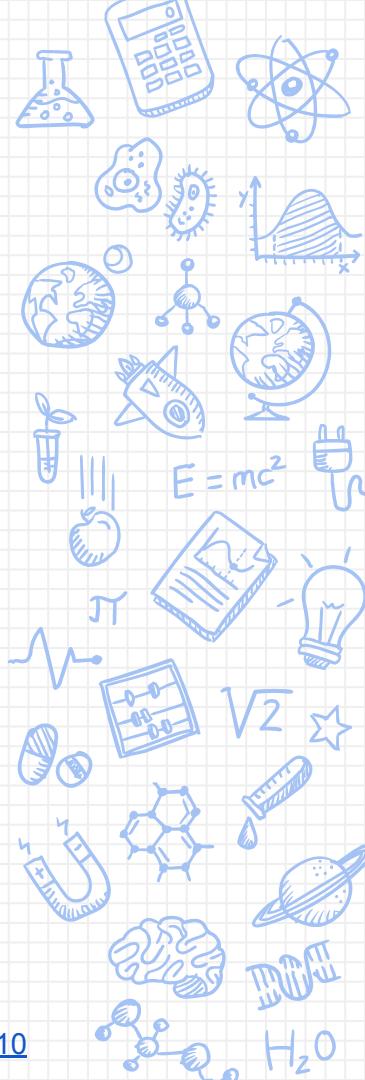


Análisis de Datos – Exploración

Consiste en entender los datos con los que contamos.



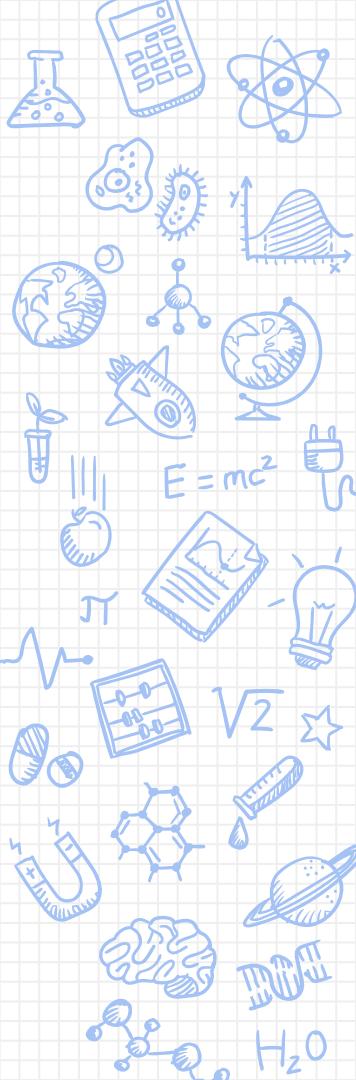
Fuente: <https://medium.com/bahmni-blog/introduction-to-exploratory-data-analysis-of-bahmni-using-r-6c186fd6f010>



Análisis de Datos – Interpretación

Capturar el significado real de los datos en el negocio.

- Reportes
 - Insights
 - Visualización de datos
 - Simulación - Rutina en EEUU
 - Gráficos - Marcas
 - Infografía - Redes Sociales



Tools



matplotlib



R Studio®

+ a b l e a u®
S O F T W A R E

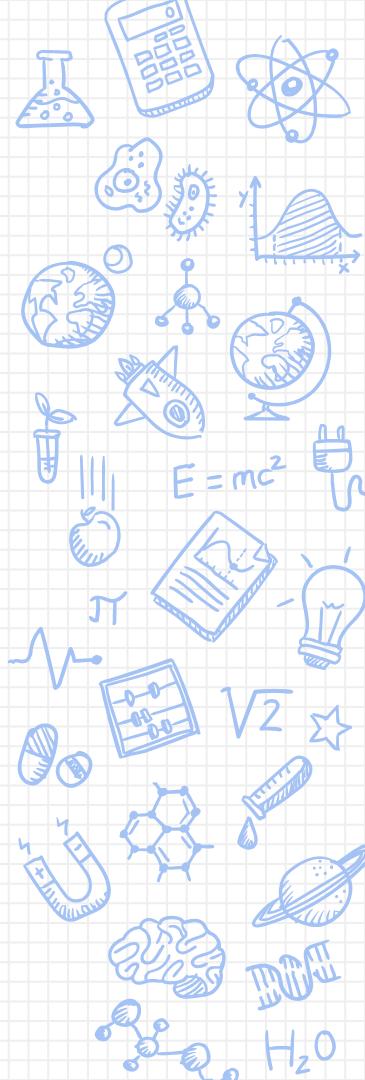
plotly

pandas

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

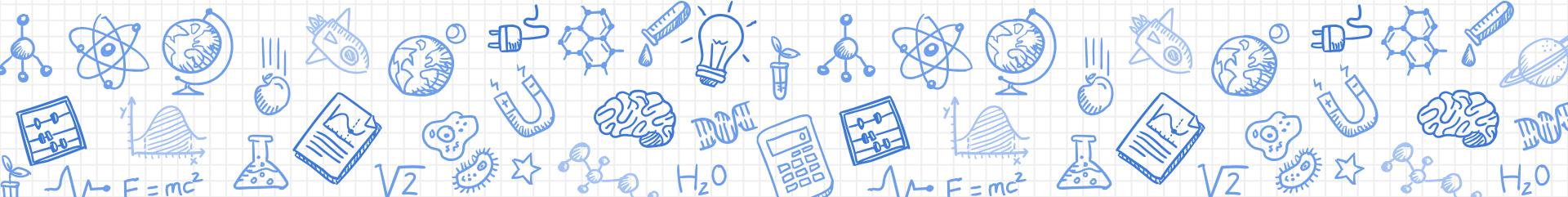


co

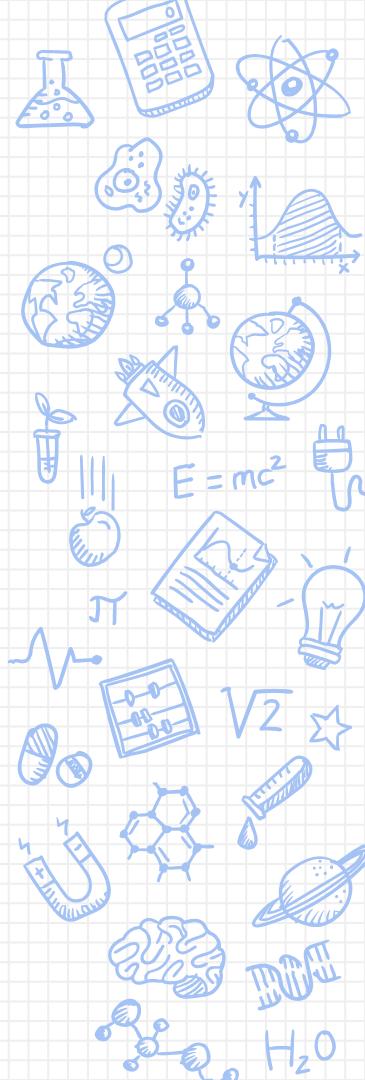
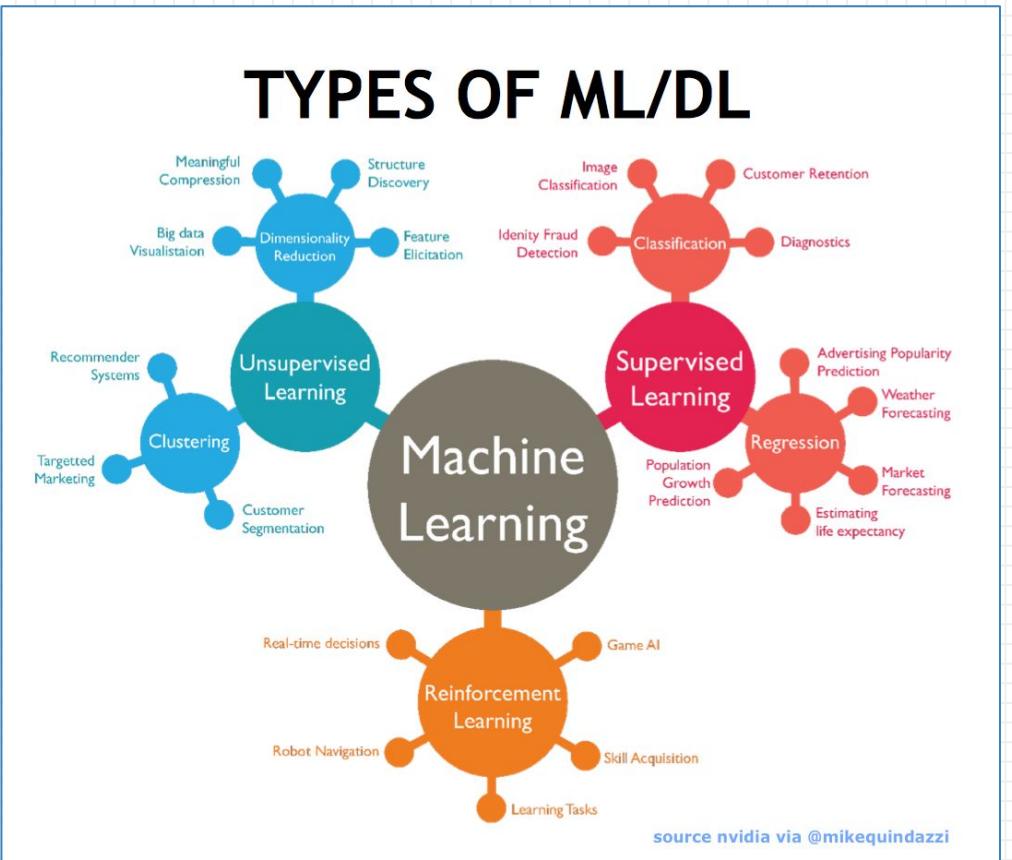


Modelado

Creación de modelos de Machine Learning



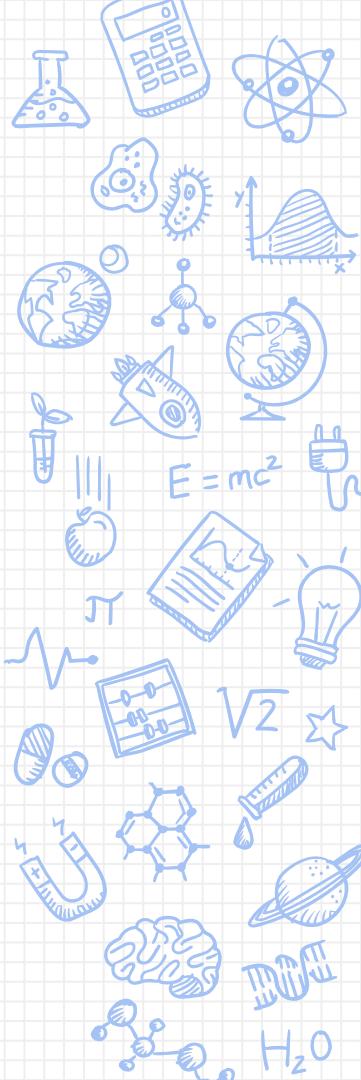
Modelado – Machine Learning Types



Modelado – Aprendizaje Supervisado

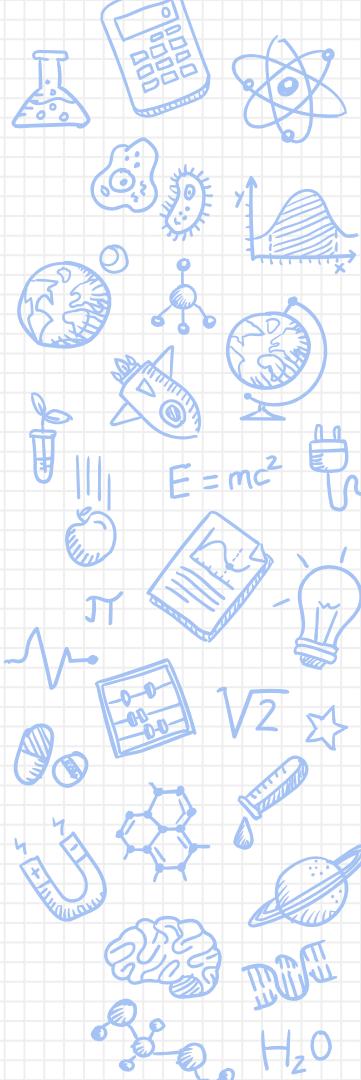
El modelo aprende en base a ejemplos que fueron previamente etiquetados por humanos.

- Ejemplos / Data points.
 - Entrenamiento.
 - Features.
 - Target.
 - Predicción.



Modelado – Clasificación

Si el target es discreto, hablamos de un modelo de clasificación.



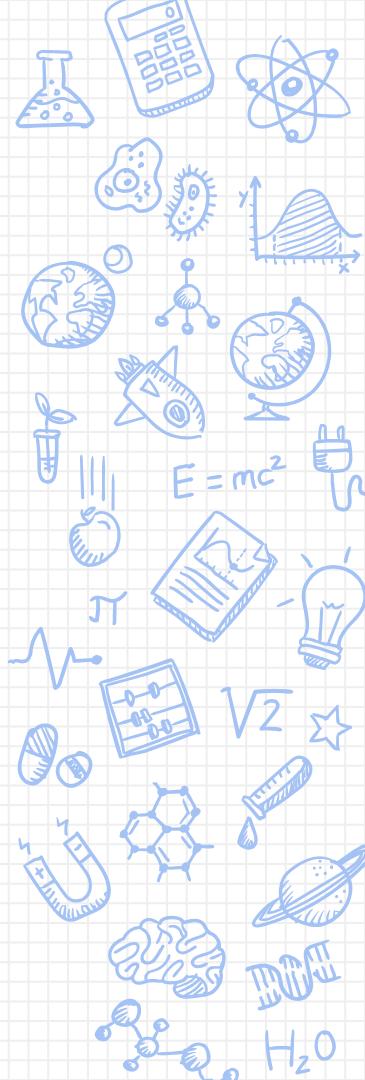
Modelado – Ejemplo de Clasificación

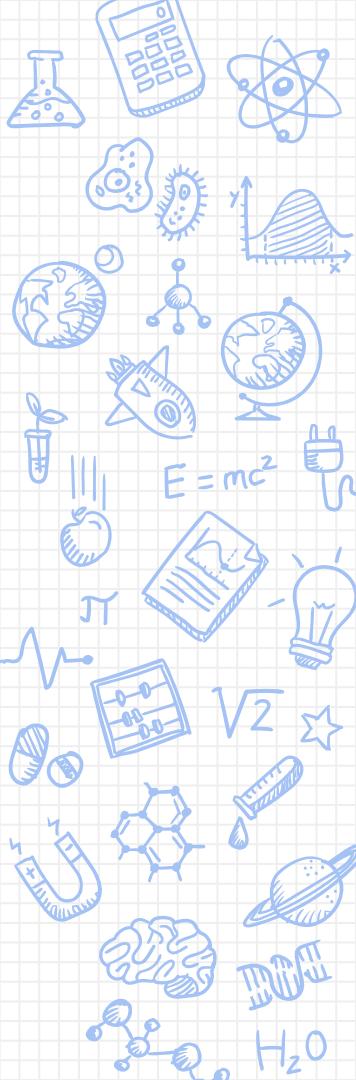
Features

Target

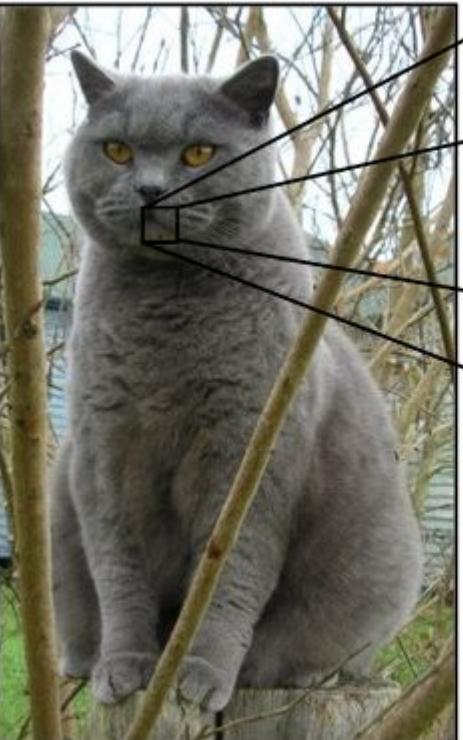
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0

Fuente: <https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/diabetes.csv>





Modelado - Ejemplo de Clasificación



What the computer sees

image classification

82% cat
15% dog
2% hat
1% mug

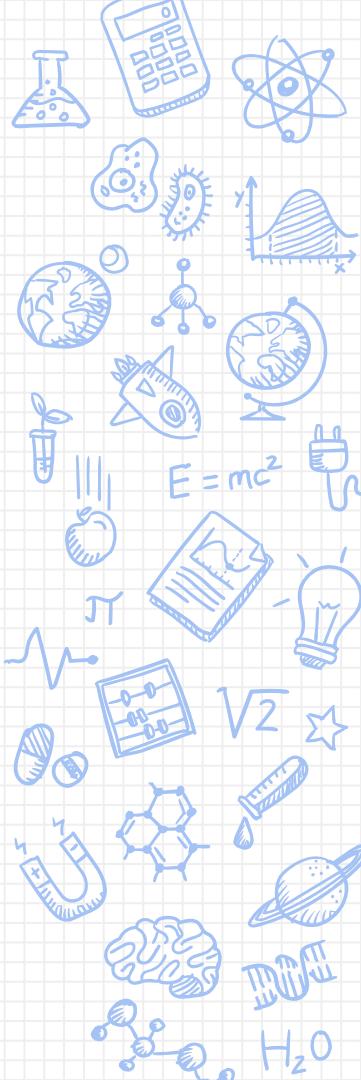
Fuente:

<https://medium.com/the-data-experience/building-a-data-pipeline-from-scratch-32b712cfb1db>

Modelado – Ejemplo de Clasificación

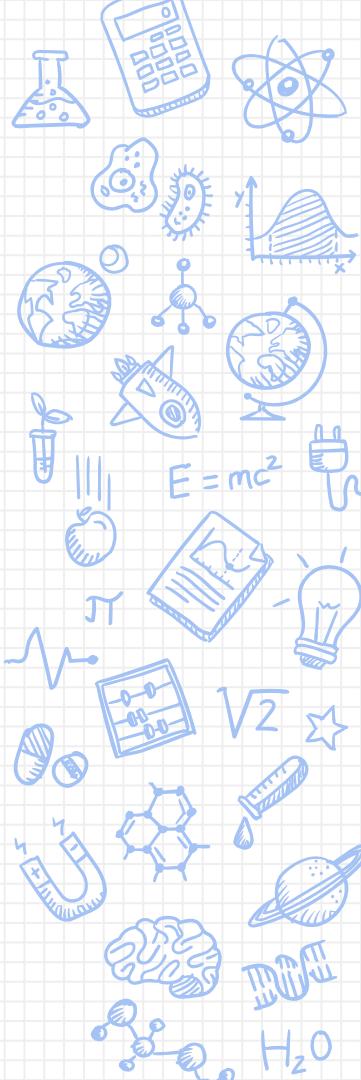
Features	Target
v2	v1
Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat... Ok lar... Joking wif u oni...	ham
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry q... U dun say so early hor... U c already then say...	ham
Nah I don't think he goes to usf, he lives around here though	ham
FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? T... Even my brother is not like to speak with me. They treat me like aids patient.	spam
As per your request 'Melle Melle (Oru Minnaminunginte Nurunqu Vettam)' has been set as your callertune	ham
WINNER!! As a valued network customer you have been selected to receivea ♣900 prize reward! To cla...	spam
Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for 1...	ham
I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough to...	spam
SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 0...	ham
URGENT! You have won a 1 week FREE membership in our ♣100,000 Prize Jackpot! Txt the word: CLA...	spam
I've been searching for the right words to thank you for this breather. I promise i wont take your help for g...	spam
I HAVE A DATE ON SUNDAY WITH WILL!!	ham

Fuente: <https://www.kaggle.com/uciml/sms-spam-collection-dataset>



Modelado – Regresión

Si el target es continuo, hablamos de un modelo de Regresión.



Modelado – Ejemplo de Regresión

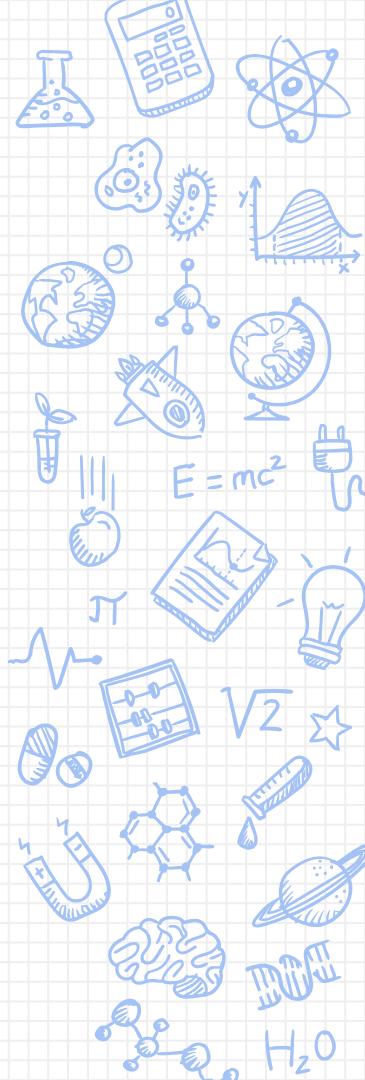
Ejemplos →

Features ↓

Target ↓

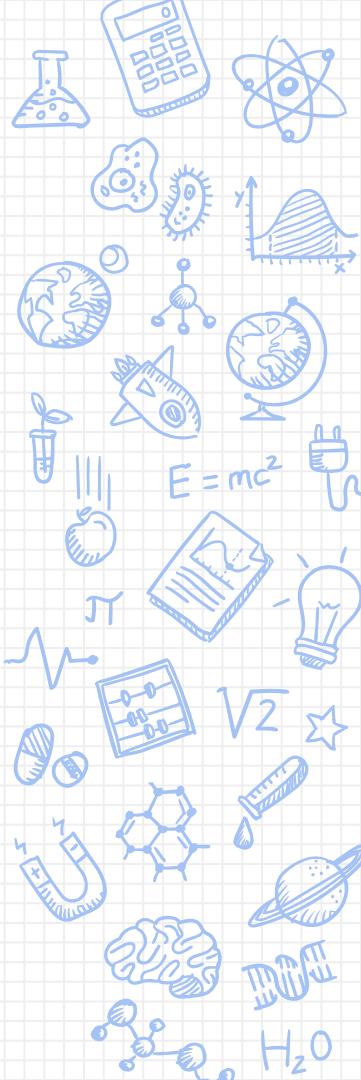
bedrooms	bathrooms	sqft_living	sqft_lot	floors	yr_built	lat	long	price
3	1	1180	5650	1	1955	47.5112	-122.257	221900
3	2	2570	7242	2	1951	47.721	-122.319	538000
2	1	770	10000	1	1933	47.7379	-122.233	180000
4	3	1960	5000	1	1965	47.5208	-122.393	604000
3	2	1680	8080	1	1987	47.6168	-122.045	510000
4	5	5420	101930	1	2001	47.6561	-122.005	1225000
3	2	1715	6819	2	1995	47.3097	-122.327	257500
3	2	1060	9711	1	1963	47.4095	-122.315	291850
3	1	1780	7470	1	1960	47.5123	-122.337	229500
3	3	1890	6560	2	2003	47.3684	-122.031	323000
3	3	3560	9796	1	1965	47.6007	-122.145	662500
2	1	1160	6000	1	1942	47.69	-122.292	468000
3	1	1430	19901	1.5	1927	47.7558	-122.229	310000

Fuente: <https://www.kaggle.com/harlfoxem/housesalesprediction>



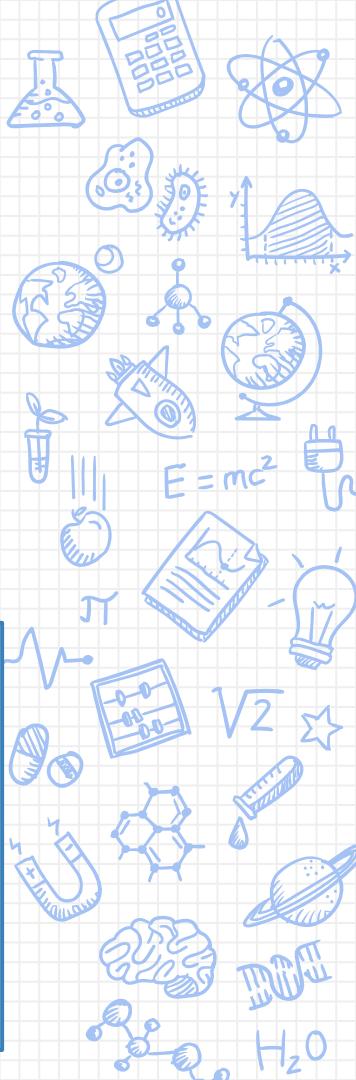
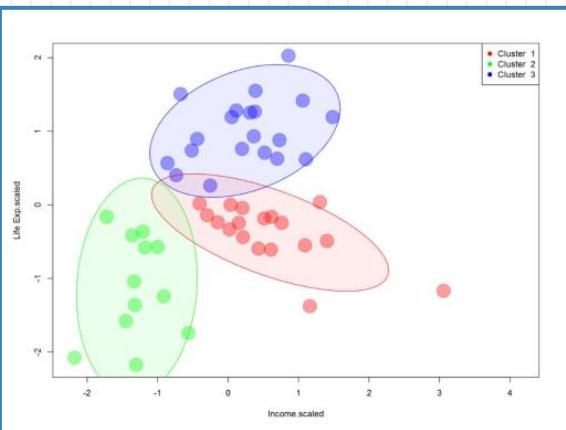
Modelado - Aprendizaje No Supervisado

El modelo aprende en base a ejemplos sin ningún tipo de etiqueta.



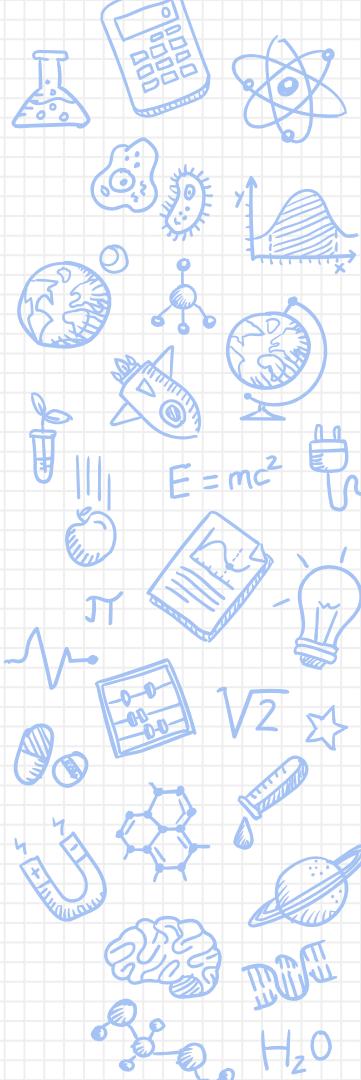
Modelado – Ejemplo de Clustering

- Segmentación de Clientes basados en sus preferencias de consumo, datos demográficos, intereses, etc.
- Clasificación en la digitalización de documentos.

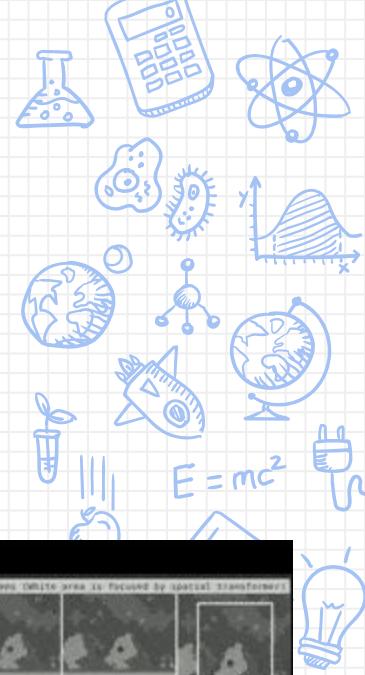
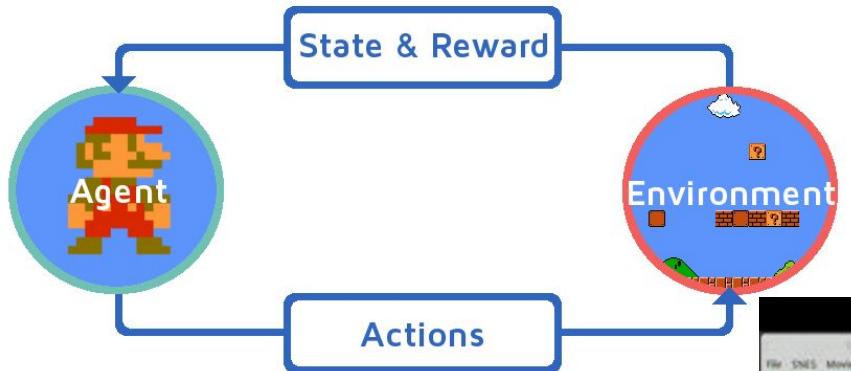


Aprendizaje por refuerzo

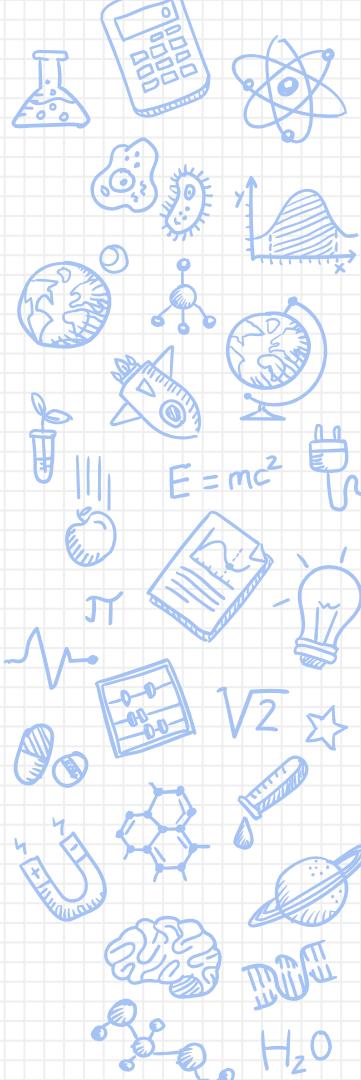
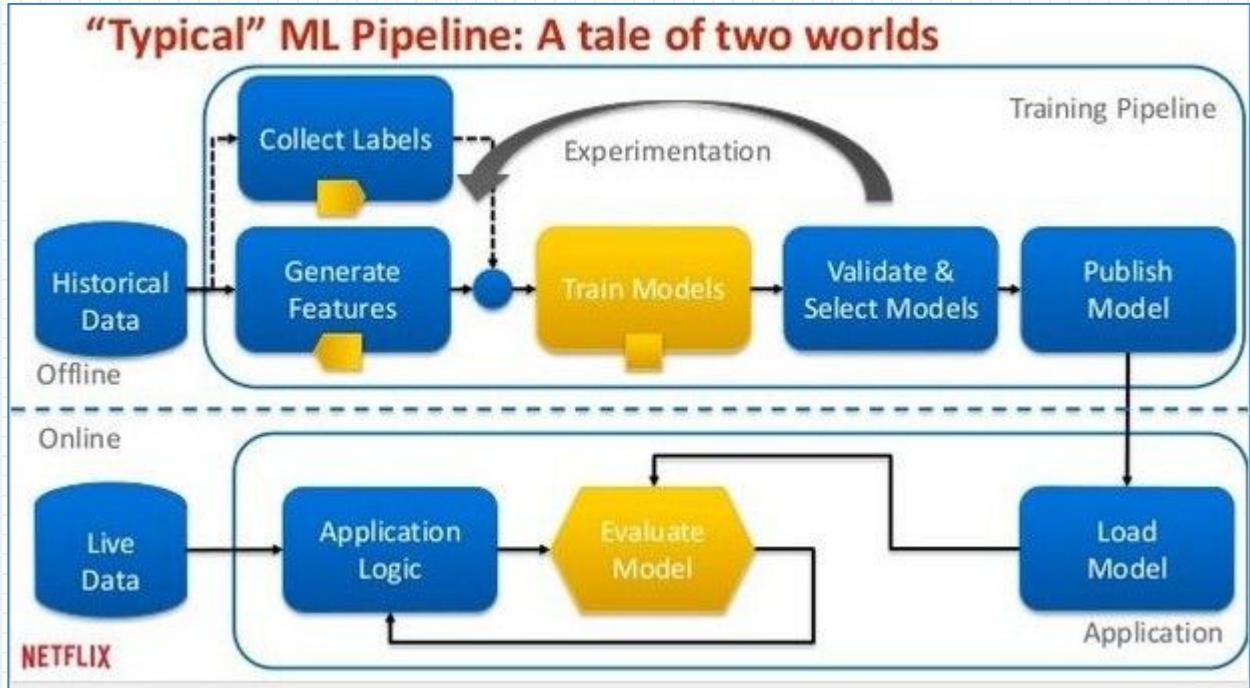
En el tercer tipo de Machine Learning, los modelos aprenden creando su propia experiencia realizando una tarea, optimizando la “recompensa” que pueda obtener ante determinada acción.



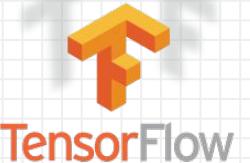
Aprendizaje por refuerzo – Ejemplos



Modelado – Ciclo de Vida de ML



Tools

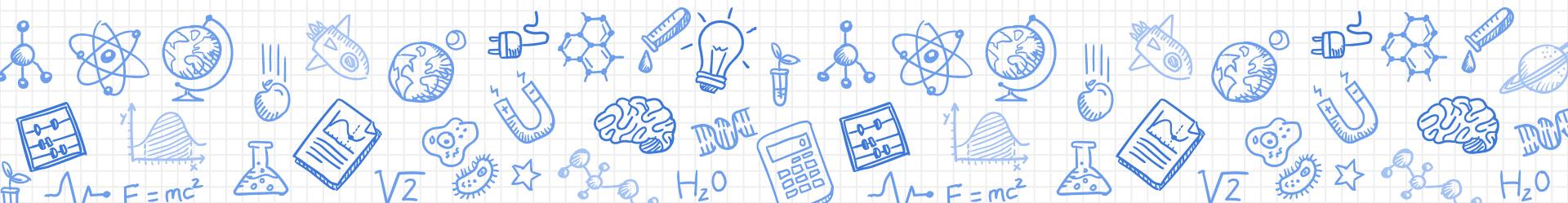


Natural Language Toolkit (NLTK)



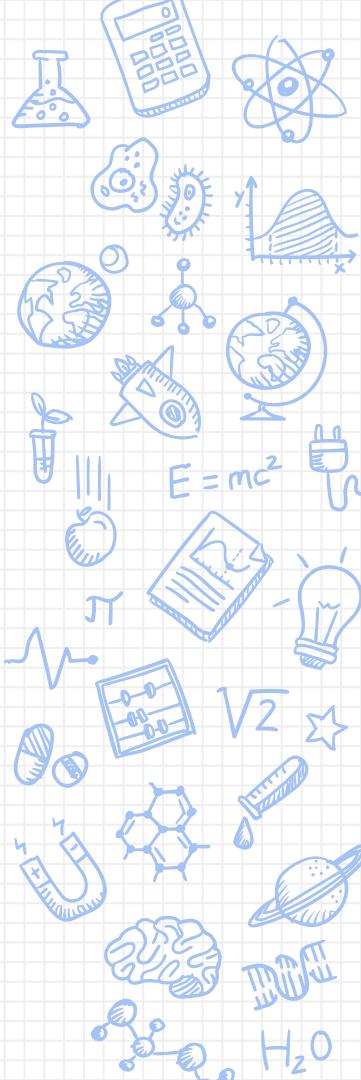
Ingeniería de Datos

Se encarga de que todo funcione
eficientemente, estable y escalable.



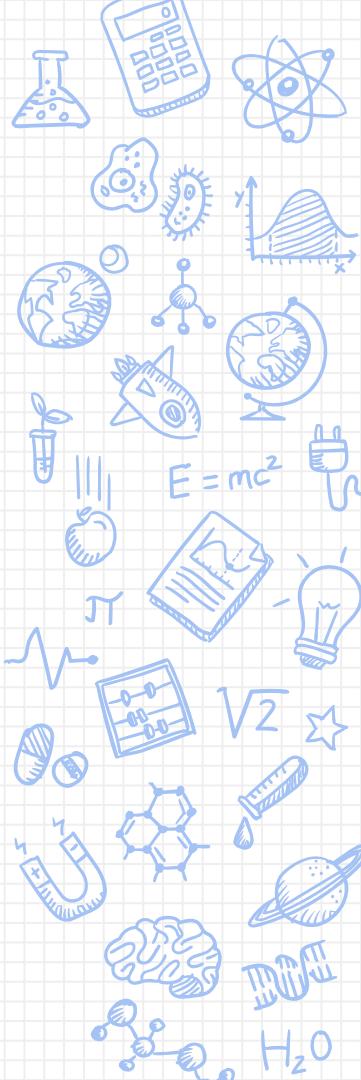
Gestión de los Datos

- Extracción de datos.
 - Transformación de datos.
 - Almacenamiento de datos.



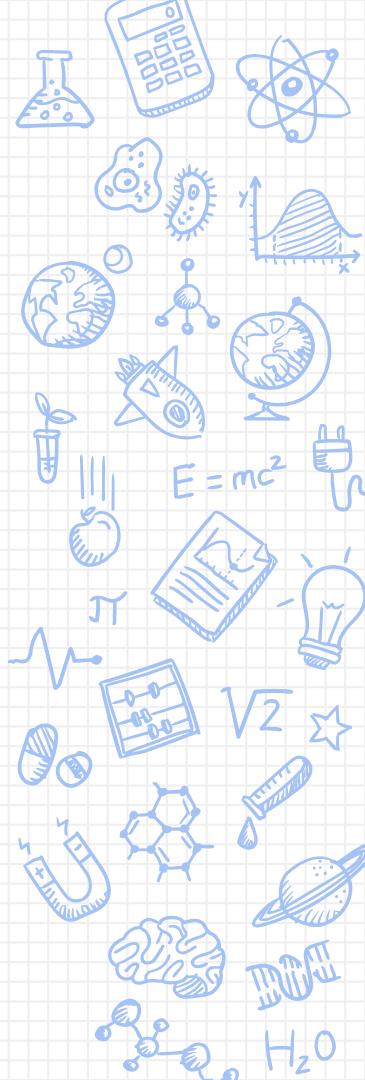
Desarrollo de Software

- Data pipelines.
 - Herramientas internas.
 - APIs.
 - Entrenamiento Automático de Modelos.
 - Monitoreo.



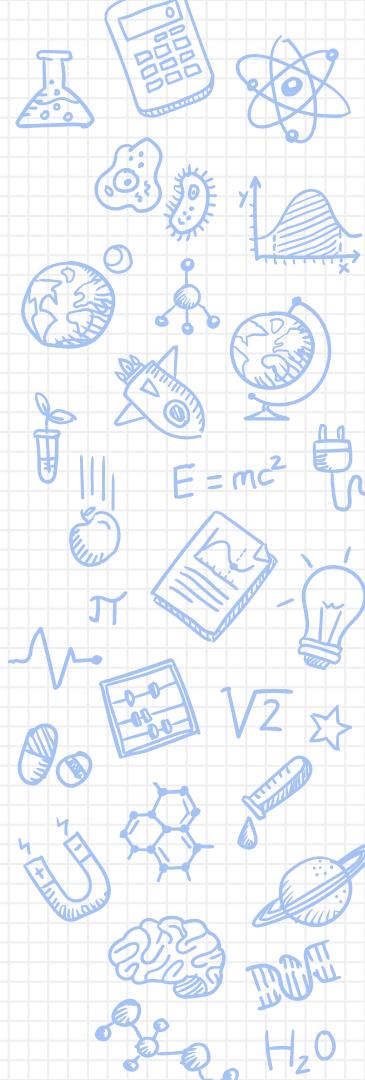
Ingeniería de Software

- Adopción de Metodologías Ágiles.
- Versionado.
- Mantenibilidad.
- Escalabilidad.
- Modularidad.

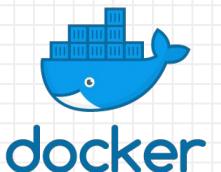
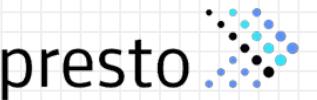


Ingeniería de Datos – Desafíos

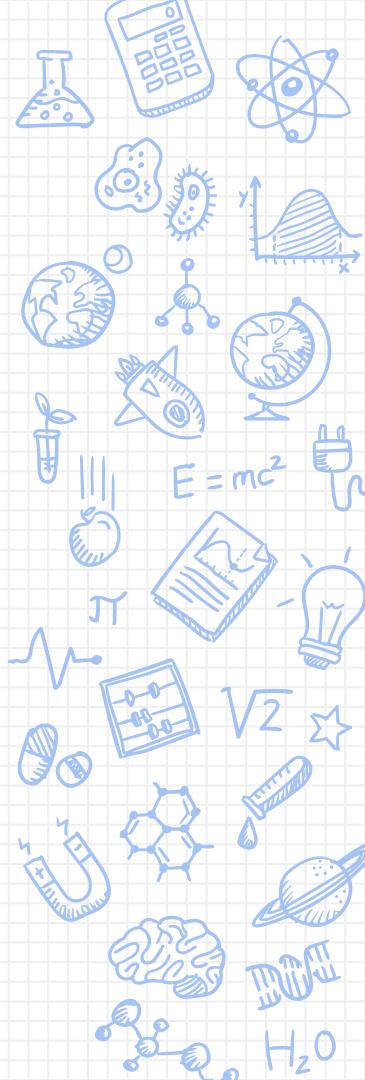
- Procesamiento Real-Time.
- Multiprocessing y Multithreading.
- Conurrencia y Asincronía.
- Tolerancia a Fallas.
- Almacenamiento Distribuido.



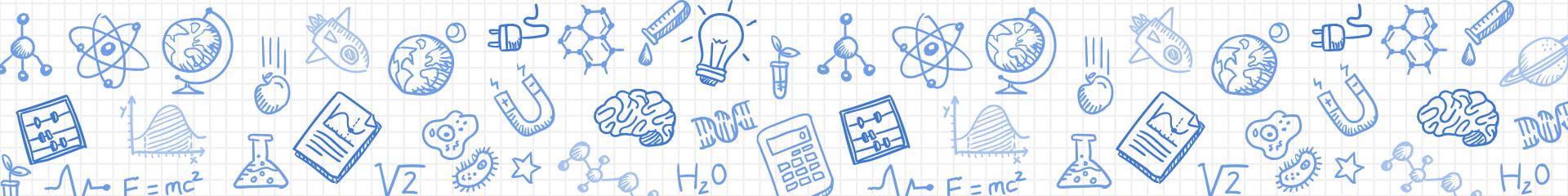
Tools



kubernetes



Skills & Jobs



7 de abril a las 20:01

Buenas! está buscando recomendaciones.

Es un tema que me interesa mucho!

Hola! Quiere saber si hay administradores de empresas que recomiendan libros o sitios web en los que se habla de negocios y administración.

Puedo ser un analista de ventas? o por ejemplo que no estudio para poder mostrar necesidades. El mismo no tiene conocimientos profundos de una disciplina como la medicina o lo que fuese en los datos que este

Buenas! Económico "sistema" cosa

Hola, me pides recomendaciones. Estoy en busca de un estudiante de sistema de información.

Buen Día a todos! Hoy les comparto un enlace

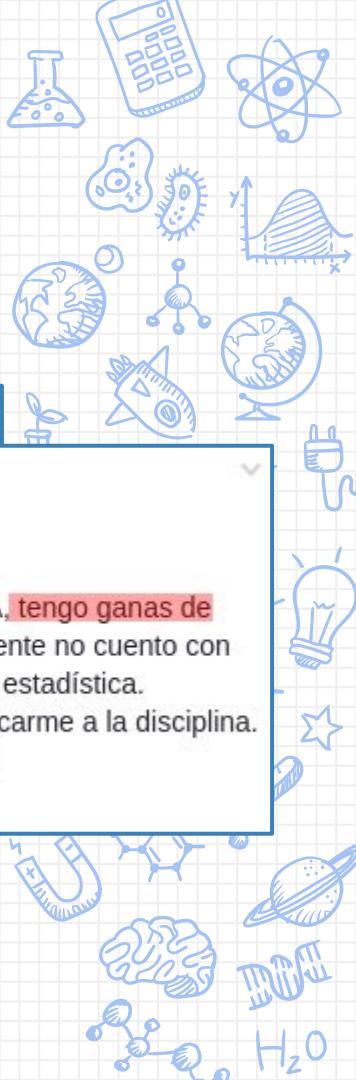
Hola gente! Hoy les comparto un enlace

11 de julio de 2018

Hola como andan?

Me recibí hace unos años en Ciencia Política en la UBA, tengo ganas de incursionar en el mundo de Data science, lamentablemente no cuento con formación en programación y tampoco tengo mucho de estadística. Necesito si alguno puede orientarme desde donde acercarme a la disciplina. Mi idea era hacer un curso en digital house o en EANT. Escucho consejos! Gracias!!

- Estadísticos.
- Informáticos.
- Contadores.
- Economistas.
- etc!



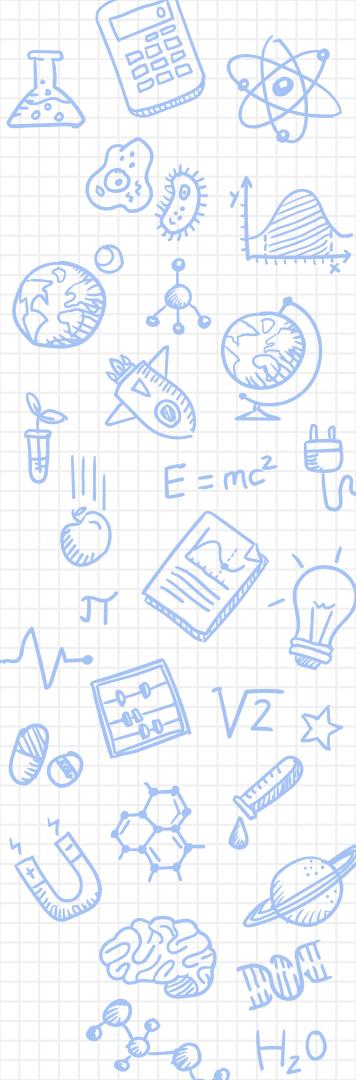
Jobs - Analista de Datos

PRINCIPALES RESPONSABILIDADES:

- Deberá colectar, procesar, y realizar análisis estadísticos sobre big data, para convertirla en descubrimientos y recomendaciones, usando técnicas de minería de datos.
- Deberá tener la habilidad de transmitir y visualizar la historia que cuenta la data.
- Investigará el comportamiento de diferentes segmentaciones de clientes para detectar oportunidades.
- Participará en la construcción y definición de los objetivos de los equipos de trabajo.
- Generará modelos predictivos con el objetivo de optimizar recursos.
- Elaborará estudios analíticos y estadísticos ad-hoc para diferentes áreas de la empresa.

REQUERIDOS:

- Nivel de Educación: Graduados en Ciencias Económicas, Sistemas o afín.
- Tiene que tener una mente curiosa, intuitiva, analítica y apasionada por descubrir.
- Proactividad y predisposición
- Conocimiento de herramientas de minería de datos como Python, R o SQL.



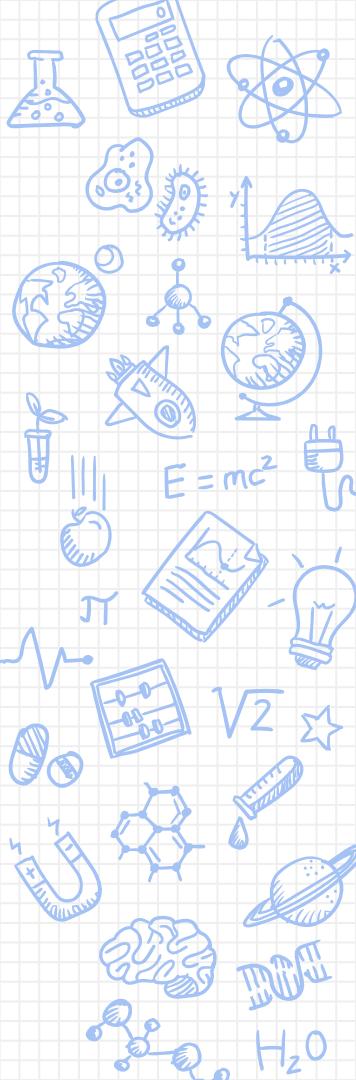
Jobs - Analista de Datos

RESPONSIBILITIES:

- Assist in the day-to-day reporting and data analyst functions to support a global performance marketing team. Covering channels including: .com performance, search data, social data, programmatic data, and offline data feeds from Salesforce / Customer feeds
- Working with various data feeds and assisting with the integration of data across various platforms
- Helping assess existing marketing reporting framework, and how this can be improved through automation and advanced data-integration techniques
- Assist the performing marketing teams with ad hoc data & insight questions about markets, channels or specific trends
- Working with the marketing channel execution leads to help identify key insights in performance, and challenge areas that need to be addressed
- Role will report into regional performance marketing leads, with dotted line to global data science and BI Director

QUALIFICATIONS:

- BA/BS, MBA or equivalent work experience in Statistics or other quantitative field
- 1-2+ years of marketing analysis, or digital marketing experience
- Extremely strong in Excel, Power BI and data visualization tools, certificates preferred.
- Familiarity with data query solutions, such as SQL, Python or R.
- Understanding of Web Analytics tools, particularly Google Analytics
- Ability to easily transition from high level strategic thinking to detailed execution



Jobs - Machine Learning

Responsibilities

- Take responsibility for ensuring that Machine Learning code, models and pipelines are deployed successfully into production, and troubleshooting issues that arise.
- Design solution architectures for applications that will use the machine learning models.
- Deploy applications to AWS's cloud leveraging the full spectrum of the AWS cloud services.
- **Automate model training and testing and deployment via machine learning continuous delivery pipelines.**
- Design and implement metrics to verify model and algorithm effectiveness.

BASIC QUALIFICATIONS

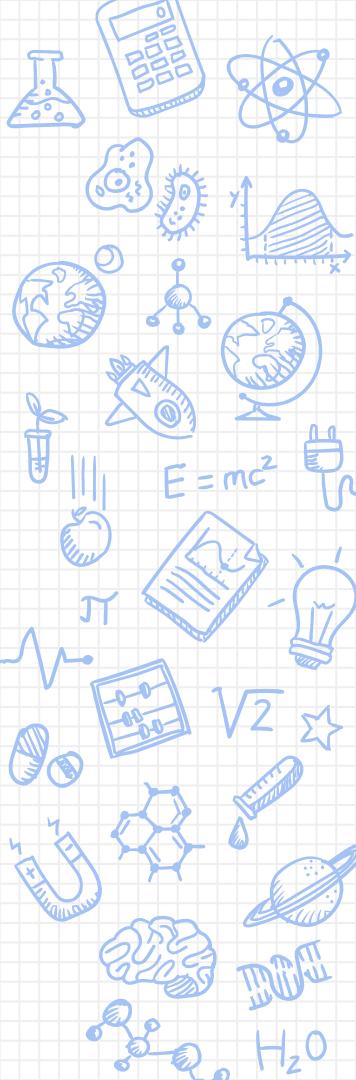
- **Master's Degree in Computer Science or Bachelor's Degree** with 5+ years related work experience
- 2-4 years of experience building production quality and **large scale deployment of applications related to natural language processing or related machine learning applications.**
- **Strong programming skills** in at least one object oriented programming language (Java, Scala, C++, Python, etc.)

PREFERRED QUALIFICATIONS

Expert familiarity with of a variety of classic and modern machine learning techniques including **deep learning, clustering, decision tree, classification, regression and neural networks.**

Solid background in algorithms, data structures, and object-oriented programming

Knowledge of mining complex data (including structure and unstructured), identifying patterns, and feature engineering.



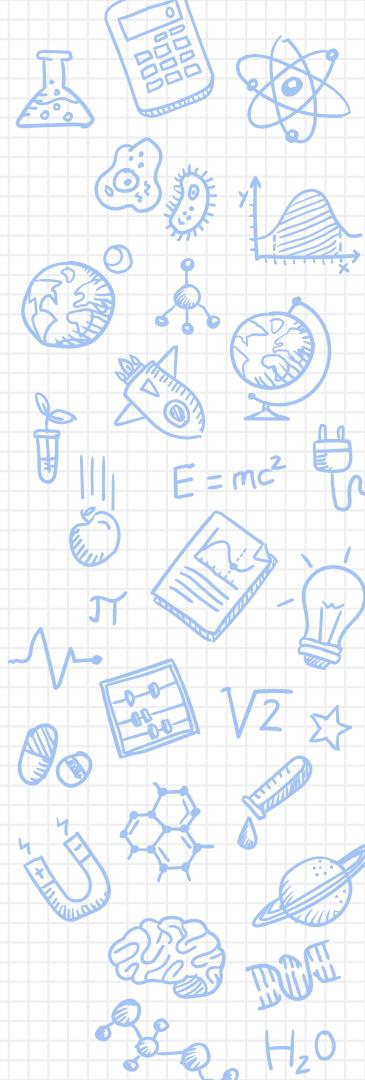
Jobs - Data Engineering

Responsibilities

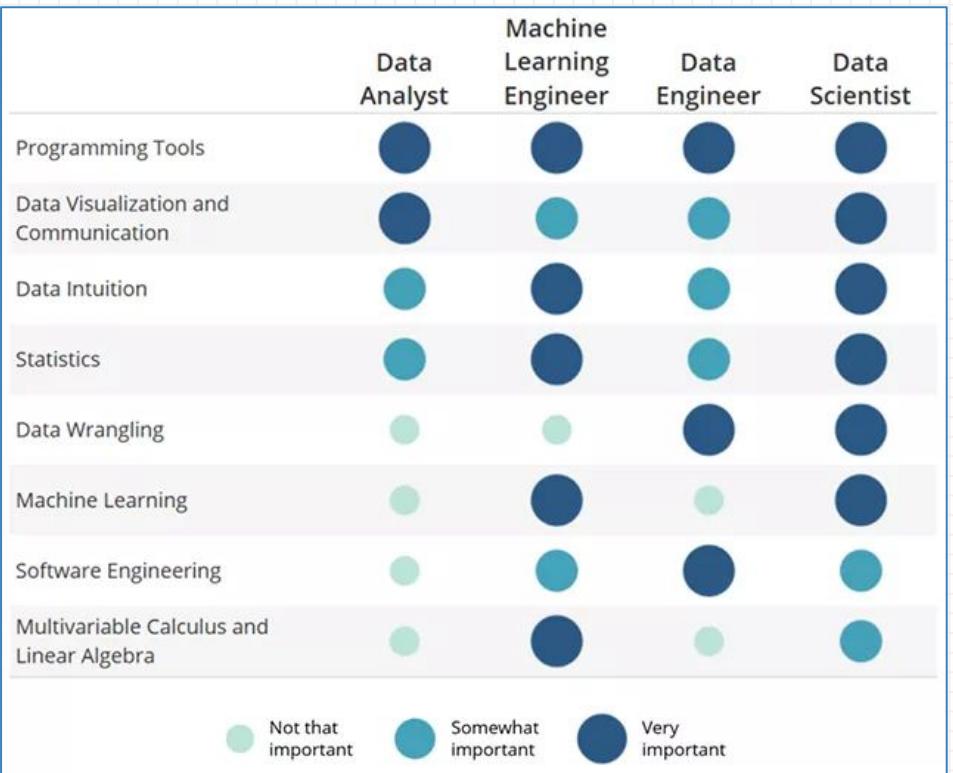
- Understand and implement mathematical approaches to solve complex business problems, utilizing available research at your disposal.
- Build machine learning applications that execute in an autonomous production environment and fully considers scalability for big data, error handling, exception conditions, etc.
- Perform unit testing and model validation of statistical routines.
- Collaborate extensively with stakeholders, program management, and software development team members to ensure that solutions meet business needs, permit valid inferences, and have functional implementability.
- Document the technical details of your work to improve collaboration with other team members.
- Guide and mentor other modelers in best practices regarding methodology (prototype formulation and testing, data mining and analysis), as well as implementation (proper documentation, source control, and coding practices).

Qualifications

- Degree in computer science, software engineering or applied mathematics
- Comprehensive understanding of quantitative methods and extreme attention to detail.
- Proficiency and experience in some data science technologies like Python, R, Scala, Spark.
- Experience applying machine learning / data science to solve a real life problem and validating it in a productive environment.



Skills



Fuente: <https://blog.udacity.com/2014/11/data-science-job-skills.html>

H₂O



Gracias!

¿Preguntas?

<https://bit.ly/2P4vQuE>

