
H2O.ai

# Ciencia de datos y ML con Python

## Programando sin morir en el intento

**Favio Vázquez**

Científico de datos / Solution Engineer
H2O.ai

@FavioVaz


# ¡Bienvenidos!

**Favio Vázquez - Científico de datos**

- Data Scientist / Solution Engineer @ H2O.ai
- Profesor de ciencia  de datos @ UNAM
- Coordinador diplomados de ciencia de datos @ UNAM
- Fundador y CEO @ Closter
- LinkedIn Top Voice en Data Analytics
- Top 100 educador en 2021 (GFEL)


H2O.ai

# Contenido

Motivación

Inteligencia artificial

Machine Learning

Ciencia de datos

Ciclo de vida de un proyecto de ciencia de datos

Modelos de regresión

Modelos de clasificación

El lenguaje de programación Python

Taller práctico


H2O.ai

# Motivación

**“Si he visto más lejos ha sido subido a hombros de Gigantes.”** - Isaac Newton

La alfabetización en datos es vital

# La era de los datos es la era de la inteligencia artificial

## Innovación

Las compañías más importantes, valiosas e innovadoras del mundo se definen como compañías de datos, como compañías tecnológicas. La mayoría de ellas comenzaron analizar datos y a utilizar inteligencia artificial hace varias décadas, por lo que está comprobado que es una actividad esencial en cualquier compañía, no es opcional.

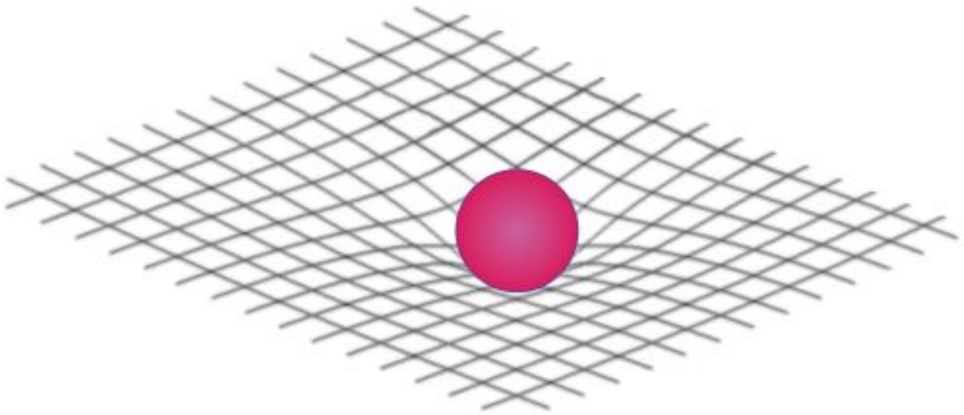

## Impacto y aceleración

Para una transformación digital exitosa, el corazón de ella son los datos y el uso que le damos a ellos. Estamos en un momento en que se está democratizando el uso de la inteligencia artificial, y la ciencia de datos es el motor que la lleva adelante. Todos los proyectos y las compañías deben transformarse en “data-driven”, hay que escuchar a los datos.

# Inteligencia Artificial (IA)

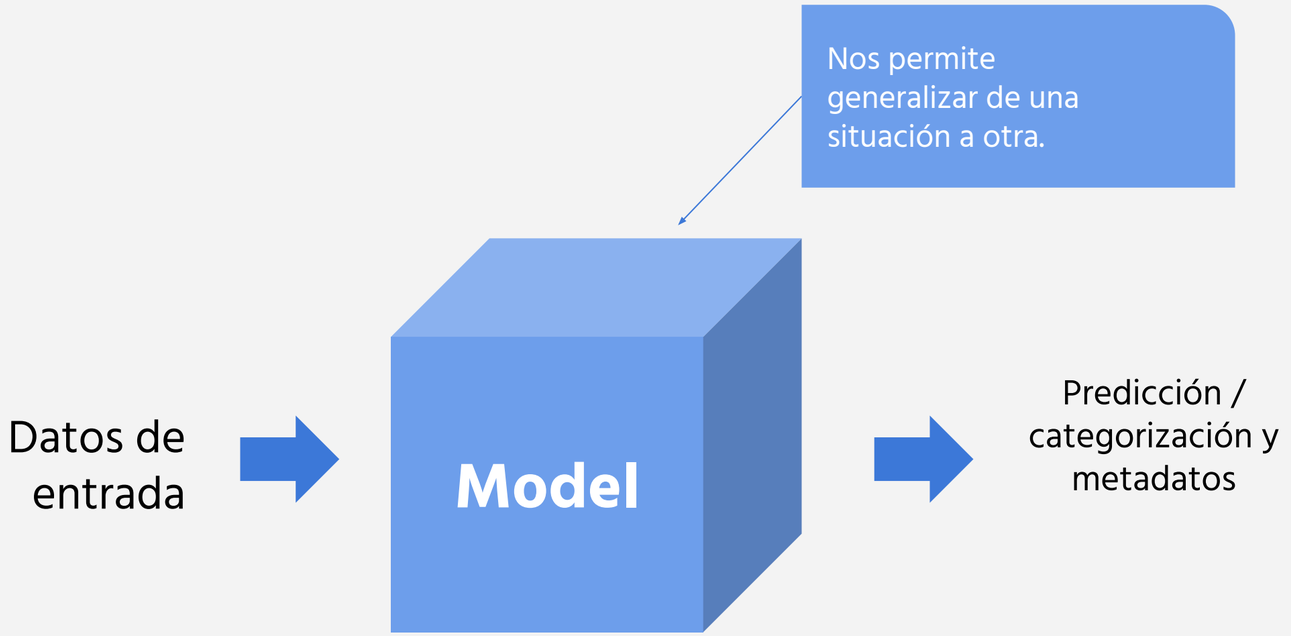

Procesos automatizados para emular las habilidades y capacidades humanas con la computación.

# Modelo

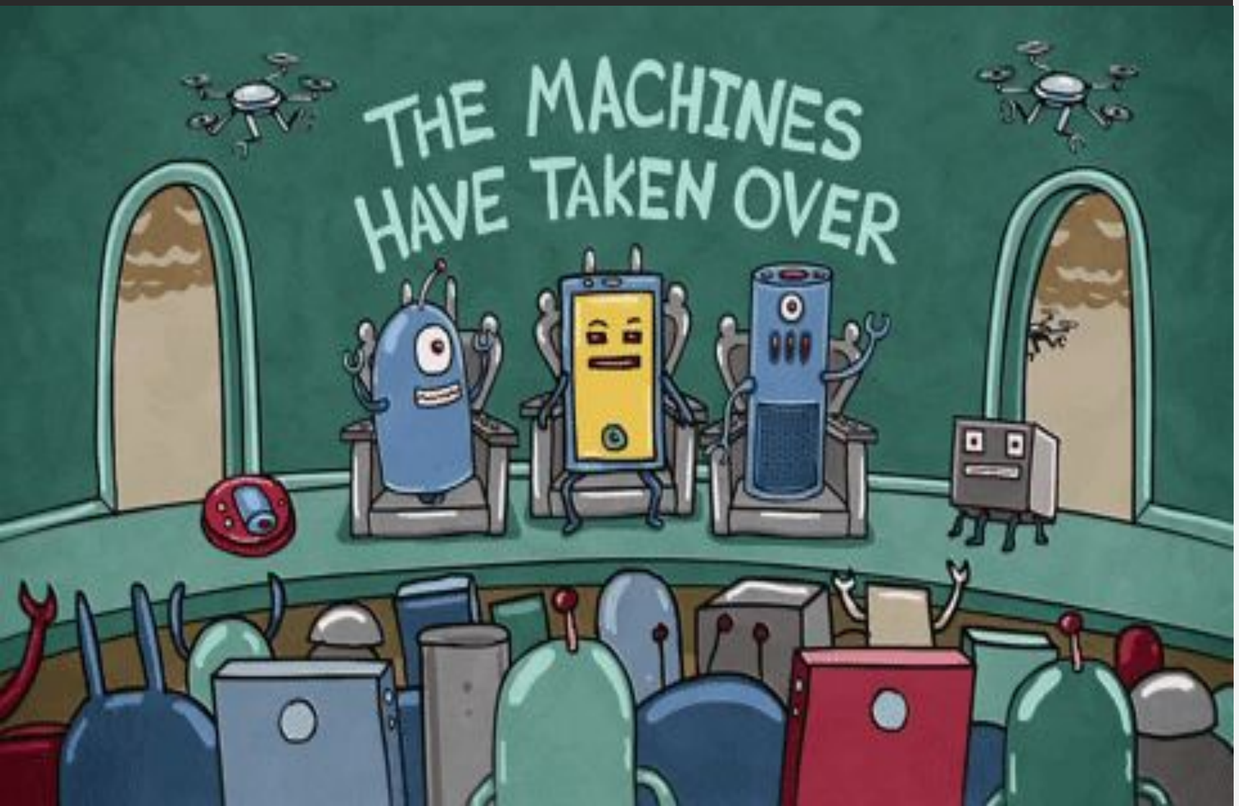


$$G_{\mu\nu}+\Lambda g_{\mu\nu}=\frac{8\pi G}{c^4}T_{\mu\nu}$$

Abstracción de la realidad para comprender un proceso o fenómeno utilizando herramientas matemáticas.

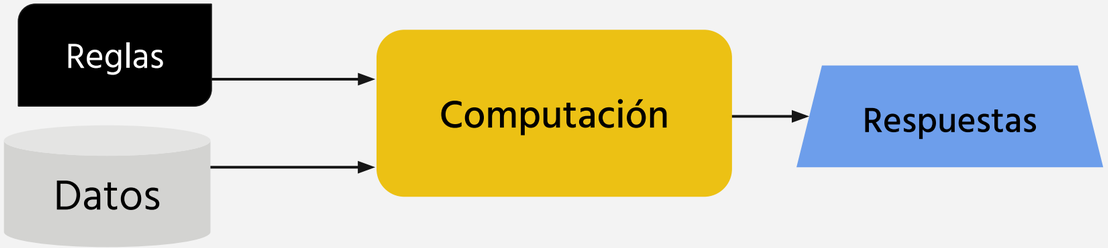


# Machine Learning

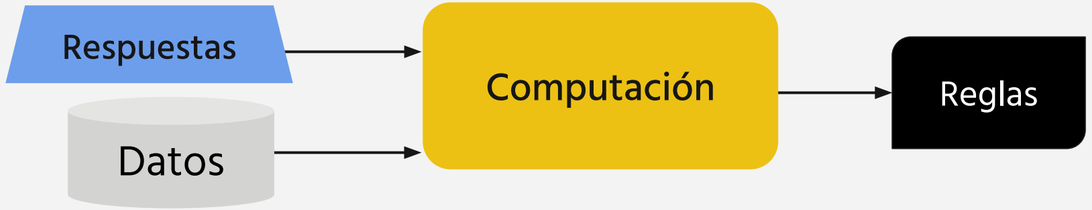



Descubrimiento de patrones a partir de datos para predecir el funcionamiento de un proceso o fenómeno utilizando algoritmos que no están programados explícitamente.

**Programación Tradicional**

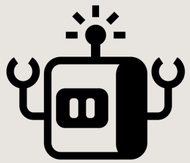



**Machine Learning**



# Tipos de Machine Learning

**Machine Learning**

| Supervisado | No Supervisado | De Refuerzo |
| --- | --- | --- |
|  |  |  |
| impulsado por tareas (predecir el siguiente valor) | impulsado por datos (identificar clústeres) | Aprender de los errores |

# EL PAPEL DE LA CIENCIA DE DATOS

La ciencia de datos es el mediador en el camino de usar la IA para impactar en negocios.


Tal vez el nombre "ciencia de datos" no sea muy cómodo para todos, pero intentaré demostrar que podemos sacarle provecho por ahora.

# Ciclo de vida de un proyecto de ciencia de datos

Recopilación de datos → Ingeniería de características → Entrenamiento de modelos → Evaluación de modelos → Operacionalizar modelos

**Reunir y preparar datos.** Este es el primer paso en cualquier proyecto de aprendizaje automático. Independientemente de la forma en que se presenten, la calidad y la integridad de los datos es un componente clave para el aprendizaje automático. Basura (datos) de entrada = basura (datos) de salida.

**Transformaciones a las variables predictoras** que pueden ser útiles en el modelado. Los ejemplos incluyen transformaciones de Box-cox, polinomios de mayor grado de variables existentes o simplemente crear algo como la columna "día de la semana" a partir de una variable temporal.

**Usar algoritmos para ajustar modelos de ML.** Experimentar con diferentes familias de algoritmos que puedan ser apropiados para el problema en cuestión. Ajustar los hiperparámetros del modelo y combine modelos que funcionen bien.

**Decida los criterios de éxito y mida el rendimiento del modelo en datos que no ha visto.** Desarrollar estrategias de evaluación y utilizar la validación cruzada para medir el rendimiento en una muestra de datos que el modelo no ha visto antes.

**Desplegar el modelo de ML.** Tome su modelo y alójelo en un entorno dedicado para que pueda hacer predicciones u obtener datos adicionales. Supervise la posible "desviación" entre el entrenamiento original y los nuevos datos

# Aprendizaje supervisado

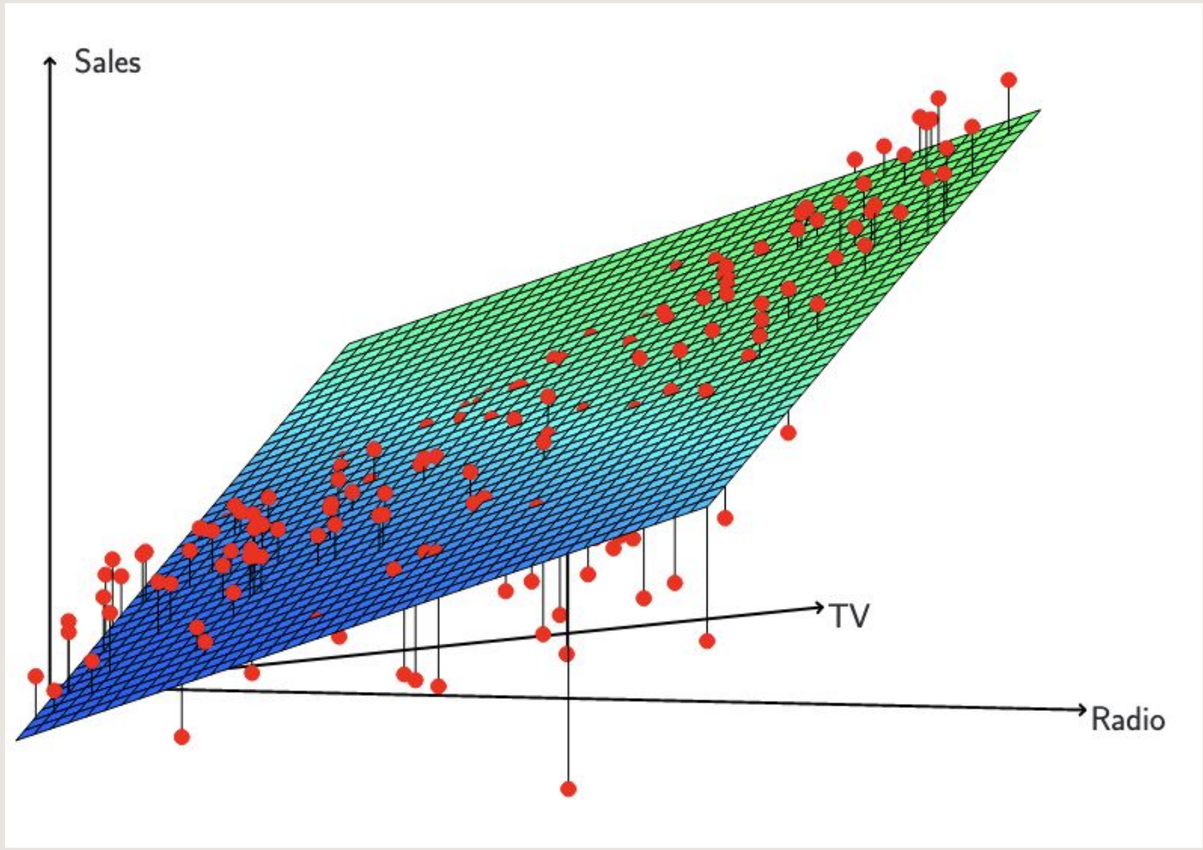
**Algoritmos:**

- Generalized Linear Model (GLM)
- Generalized Additive Models (GAM)
- Gradient Boosting Machine (GBM)
- Random Forest
- Neural Network

**Métricas de error (“Scorers”):**

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percent Error (MAPE)

**Regresión - Variable objetivo continua**




Fuente: [Introduction to Statistical Learning]

# Aprendizaje supervisado

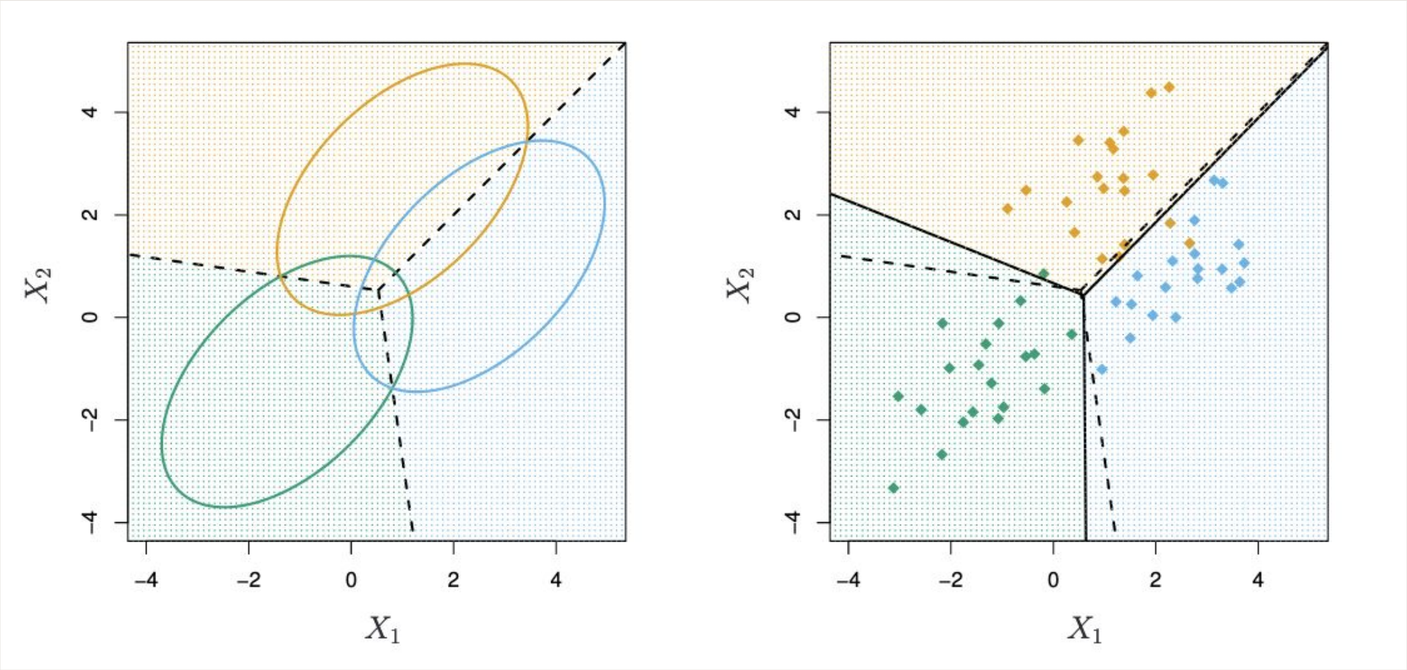
## Algoritmos:

- Generalized Linear Model (GLM)
- Generalized Additive Models (GAM)
- Gradient Boosting Machine (GBM)
- Random Forest
- Neural Network

## Métricas de error (“Scorers”):

- Area Under ROC (AUC)
- Area Under PR Curve (AUCPR)
- F1 Score
- Accuracy

## Clasificación - Variable objetivo categórica




Fuente: [Introduction to Statistical Learning]

R también es un súper lenguaje para CD

# El lenguaje de la ciencia de datos: Python

**01**
**Fácil**

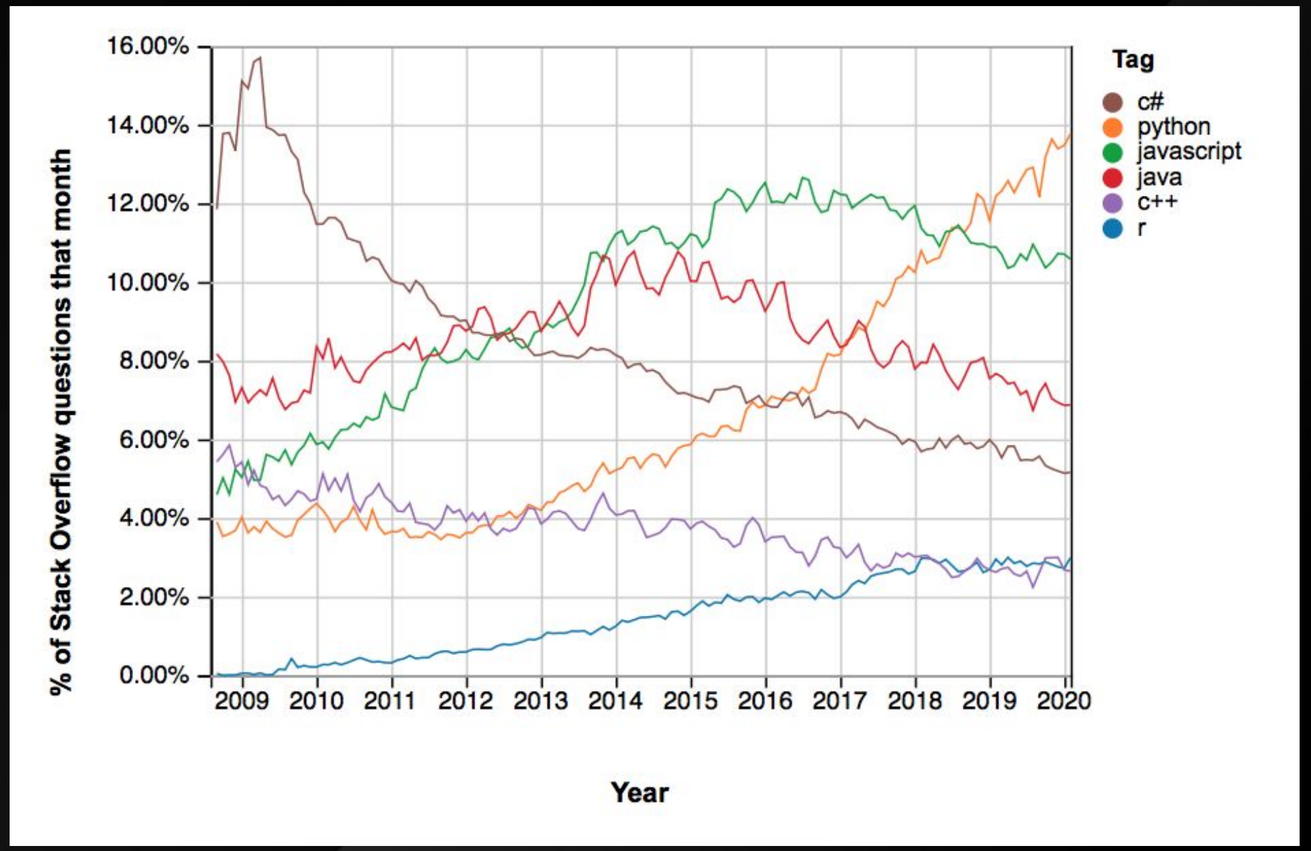
Aprender Python es una tarea fácil si lo comparamos con otros lenguajes de programación.

**02**
**Robusto**

Python es un lenguaje completo, con una comunidad activa y miles de librerías que hacen muy sencillo hacer cualquier tarea con él.

**03**
**Listo para Producción**

Es sencillo tomar un modelo de Python y ponerlo en producción, el mundo de la ingeniería de datos a nivel empresarial depende en gran parte de este lenguaje.




H2O.ai

# Thank you

**Contacto**

Favio Vázquez
Data Scientist @ H2O.ai
@ favio.vazquez@h2o.ai
in https://www.linkedin.com/in/faviovazquez/