Encontramos el anillo de Sauron

"Machine intelligence is the last invention that humanity will ever need to make" - Nick Bostrom. * Filósofo, Oxford University. *

Ya conoces herramientas y conceptos clave para dar tus primeros pasos en el mundo de Data Science. Exploraste la programación con Python en un entorno particular orientado al análisis de datos (Jupyter) y aprendiste a utilizar muchas de las librerías típicas del lenguaje. También, tuviste la oportunidad de repasar conceptos útiles de estadística y de conocer algunas técnicas de preprocesamiento de datos con Pandas. Todo esto, ¿para qué? ¡Para aplicarlo al Análisis Exploratorio de Datos! Considera lo aprendido como una caja de herramientas: verás que será muy útil para usar e interpretar datos de manera correcta.

¡Has llegado a un momento crucial del aprendizaje! Es el momento de abordar una pieza clave de este rompecabezas que es Data Science: ¿cómo hacemos para que las computadoras puedan aprender de los datos?



Para empezar a hacer predicciones es necesario entender en qué consiste un

modelo matemático y para qué sirve. Según Wikipedia los modelos matemáticos son descripciones de un sistema que

usa conceptos y lenguaje matemático, que ayuda a estudiar los efectos de diferentes componentes y hacer predicciones relacionadas al comportamiento. Este proceso de agrupar información en una herramienta para predecir se llama modelado y se estructura alrededor de una idea: por qué ciertos eventos futuros podrían suceder. Una forma básica de armar un modelado sería tomar patrones y observaciones

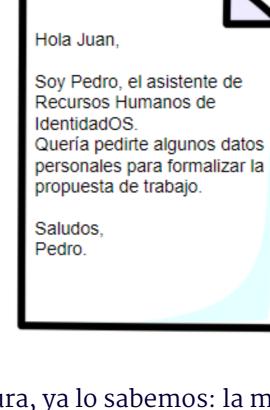
de eventos pasados que puede llevarnos a pensar que en el futuro estos componentes se repetirán. Modelos más sofisticados —como las leyes de física — contemplan y proporcionan nociones como la causalidad, que nos ayuda a entender mejor por qué las cosas suceden. Evaluar de manera precisa el desempeño de un modelo puede ser muy difícil, pero es clave a la hora de interpretar los resultados.

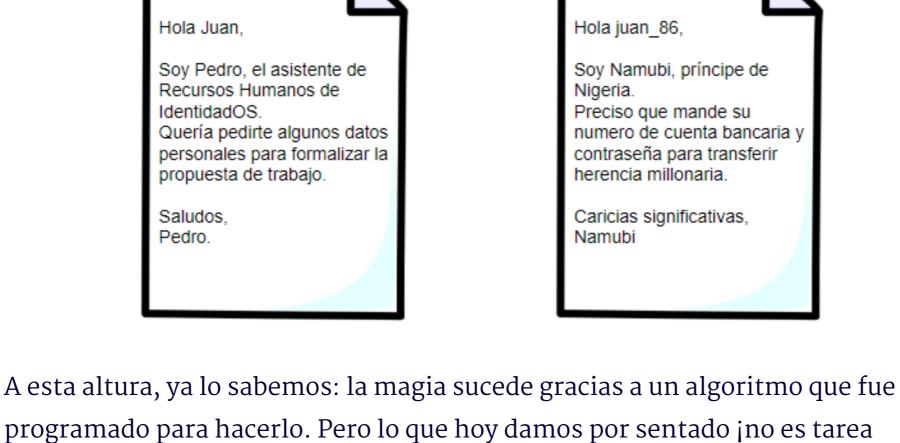
Un clásico ejemplo para entender mejor de qué se trata todo esto es la

Ahora sí ¡Hola Machine Learning!

clasificación de correo de Spam. ¿Cómo hacen las casillas para determinar qué correos destinar a esa carpeta?

¿cómo es que clasifican como Spam los mails de supuestos príncipes de alguna parte del mundo que quieren regalarnos millonadas de dinero si tan sólo le damos todos nuestros datos bancarios? ¿Y cómo es que, en cambio, no nos perdemos propuestas laborales concretas cuando llegan a nuestra bandeja de entrada?





sencilla! De hecho, necesita un nivel de procesamiento de abstracción bastante alto: comprender el texto, relacionar conceptos, buscar patrones, y todavía más. Machine Learning es una rama de la Inteligencia Artificial (IA) que se ocupa de crear este tipo de modelos a partir de datos. Supervisado vs. No Supervisado

Como ya sabes, a la hora de programar y entender los datos, es fundamental definir un tipo de abordaje. En Machine Learning, este aspecto también es

esencial. Para eso, existen dos grandes tipo de aprendizaje dentro de Machine Learning: supervisado y no supervisado. El **Aprendizaje Supervisado** (supervised learning) es aquel que permite modelar la relación entre las características medidas de los datos y alguna

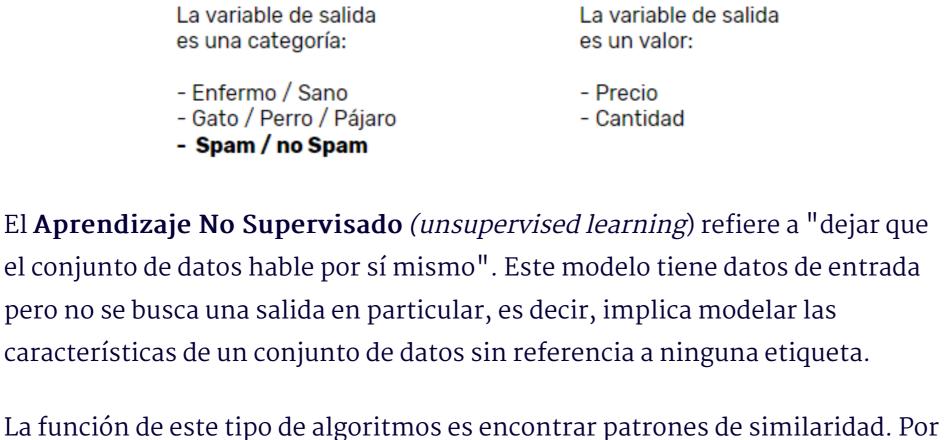
etiqueta asociada con los datos. Una vez que se determina este modelo, se puede usar para aplicar etiquetas a datos nuevos y desconocidos. De acuerdo al tipo de etiquetas que asociamos a los datos, el modelo puede realizar dos tipos de tareas:

ejemplo del spam, la tarea es de clasificación, ya que un correo puede ser o no spam (dos etiquetas).

Tareas de clasificación: las etiquetas son categorías. Por ejemplo, en el

Tareas de regresión: las etiquetas son cantidades continuas, valores. Entre las tareas de regresión, por ejemplo, podríamos determinar el valor de una propiedad o el precio de un producto.

Clasificación Regresión



eso, los modelos de Aprendizaje No Supervisado incluyen tareas como la agrupación (clustering) y la reducción de dimensionalidad. Algoritmos de reducción de dimensionalidad: buscan representaciones más concisas de los datos.

Algoritmo no Datos de Patrones Supervisado encontrados entrada **X**

(grupos)

Predecimos

Entrada X

Modelo

Algoritmos de clustering: como su nombre lo indica, identifican distintos

grupos de datos. Por ejemplo:



• Cuenta con una excelente documentación • Es Open Source ;)

• Es eficiente

Creamos

Selección de modelo

y parámetros

Modelo

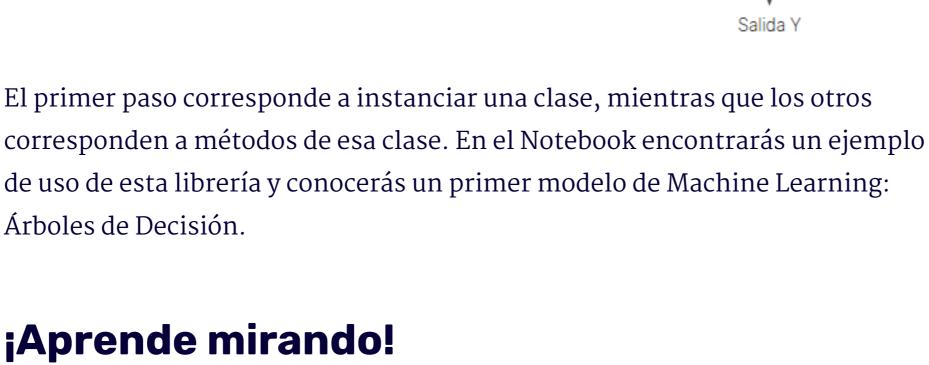
En Scikit-Learn, prácticamente todo lo que usamos son objetos y métodos de esos objetos. Su forma de trabajo se puede resumir en la siguiente figura:

Entrenamos

Training data

ΧeΥ

Modelo



Árboles de Decisión. ¡Aprende mirando!

Recién estamos en los comienzo de Machine Learning, pero probablemente te sirva ir mirando este video sobre el desarrollo de un modelo. Toca temas que todavía no vimos - algunos, de hecho, corresponden al Sprint 02 - pero te

servirán para ir embebiéndote de conceptos y prácticas. ¡No te preocupes si te pierdes con muchos de los conceptos! Pronto quedarán más claros. Capítulo 4: Machine Learning Challenge

Empieza a trabajar en el Notebook

Instala Scikit-Learn (recuerda activar el ambiente y usar conda) y trabaja en el siguiente notebook, ya sea en Colab o en tu computadora. Para el encuentro, lleva resuelta la Sección 1.



¡Prepárate para el próximo encuentro!



Notebook: DS_Bitácora_11_ML Challenge Profundiza

Lista de recursos