





Introducción a la Ciencia de Datos

Sesión 1:

Pablo Benavides - Noe Camacho - Enzo Falcon

INTRODUCCIÓN A DATA SCIENCE

SESIÓN 1





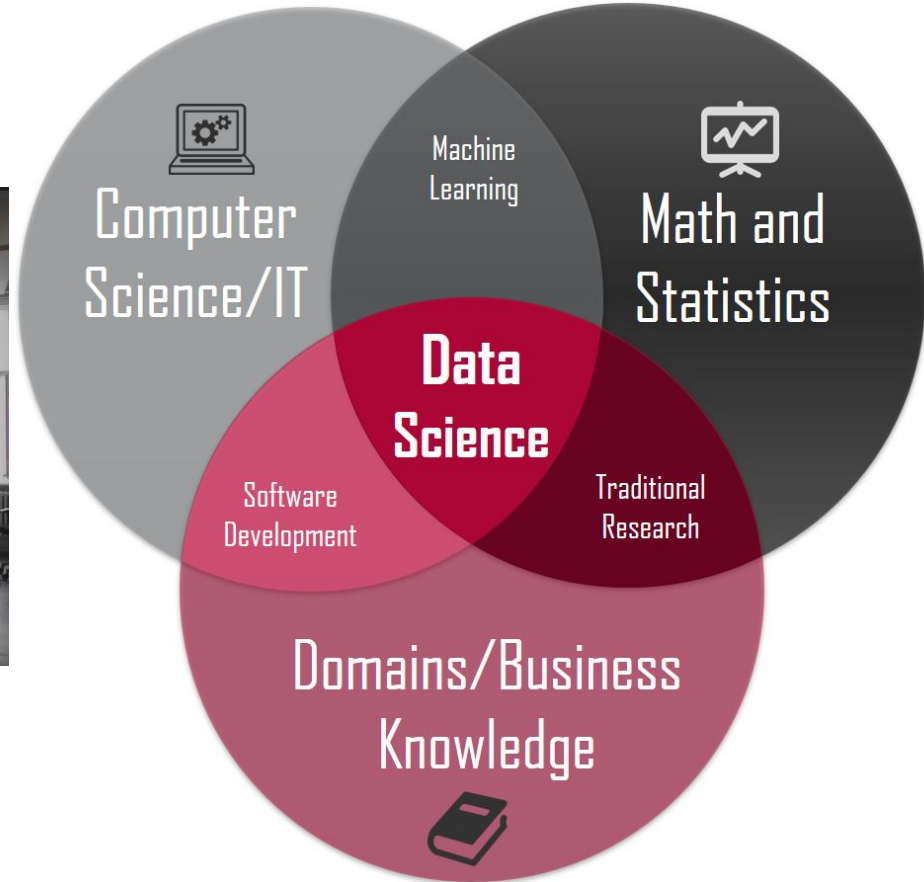
RECRUITER



TABLEAU

IS THIS A DATA SCIENCE?

¿Cuáles son las herramientas de un Científico de Datos?



¿cómo y con qué empezar?



VS.



PLATAFORMAS COLABORATIVAS

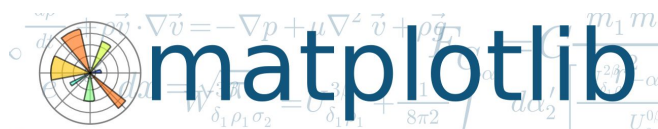
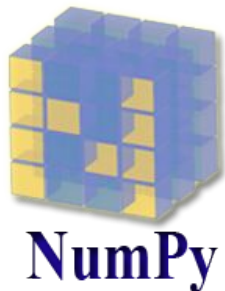


Medium

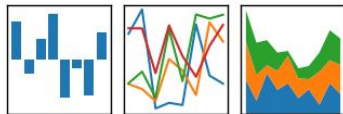
colab

Towards
Data Science

kaggle

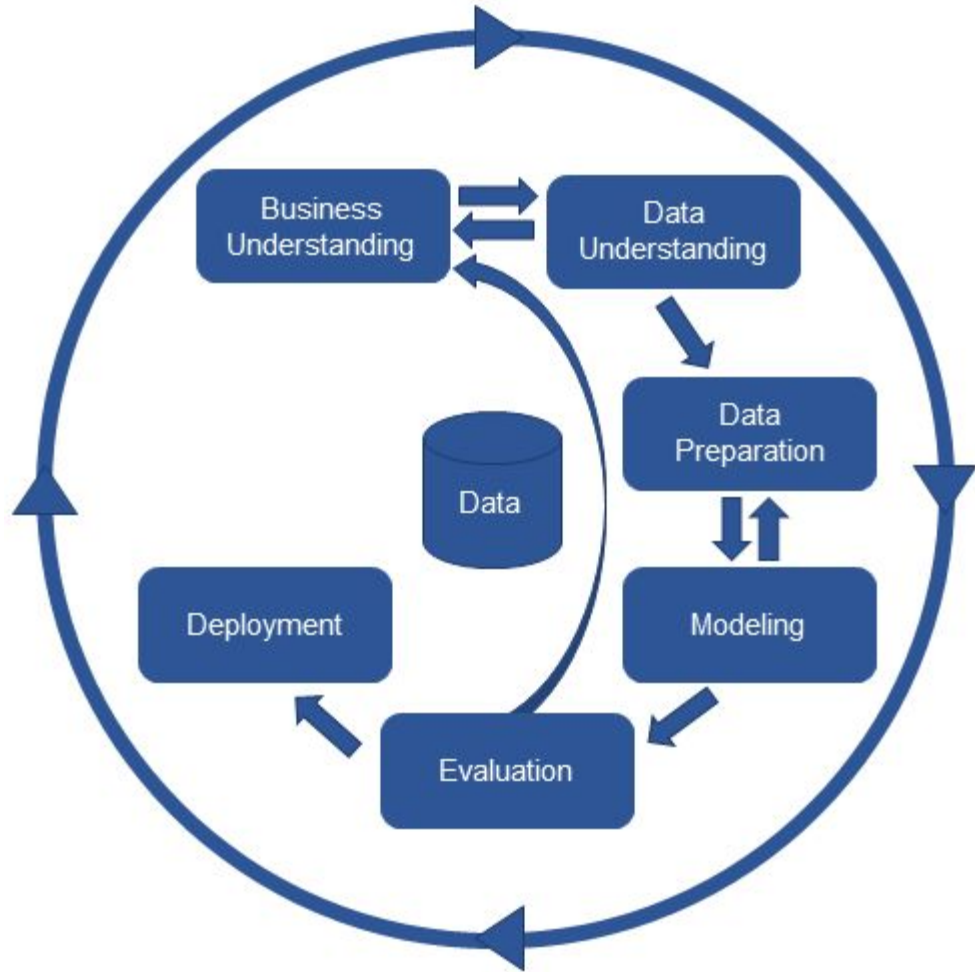


pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



METODOLOGÍA

CRISP - DM



**Identifica una
necesidad del
mercado**

1

¿Algún problema
que puedas
resolver aplicando
análisis de datos?

**Captura los
datos
necesarios y
limpia la
información**

2

¿Tienes todos los
datos que
necesitas? ¿Los
datos están
legibles?

**Selecciona y
evalúa el
modelo**

3

¿El modelo es el
mejor para los
datos que tienes?
¿El modelo es
preciso?

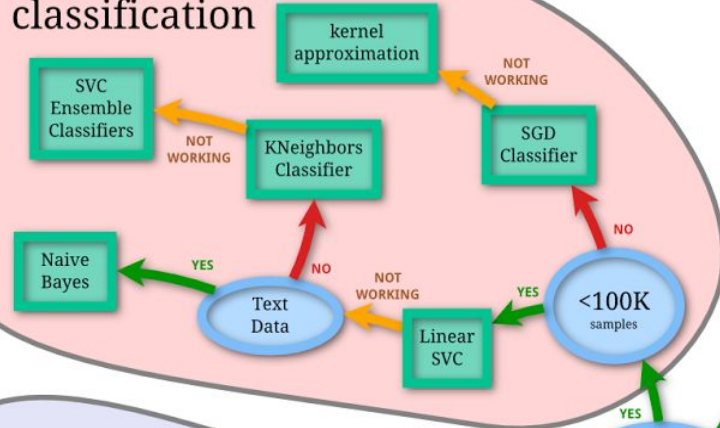
**Implementa el
modelo para
uso de tus
clientes**

4

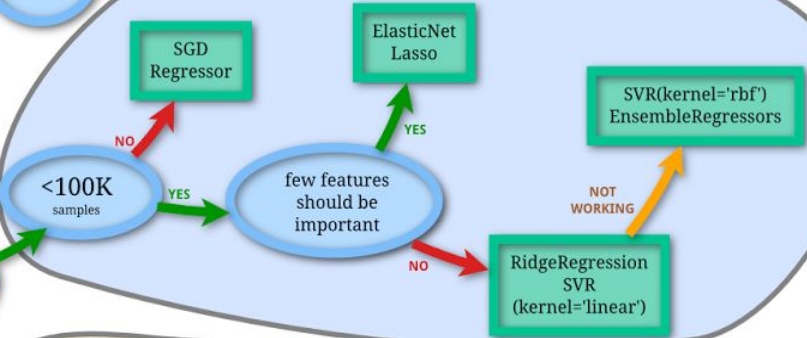
¿Tu modelo puede
salir a producción?
¿Tienes la
infraestructura?

scikit-learn algorithm cheat-sheet

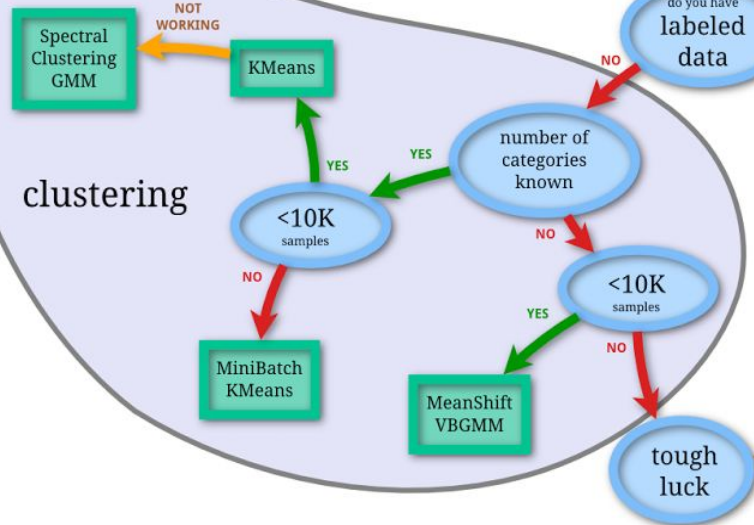
classification



regression



clustering



dimensionality reduction



accelerate your learning

Normas de clase online:

- Se brindará una tolerancia de 15 min para dar inicio a la clase, en caso de emergencia por favor comunicarse con los profesores con anticipación.
- Habrá un break de 10min cada 40-45min.
- Se podrá admitir hasta un máximo de **dos (2) inasistencias** para obtener el certificado.
- Se podrá hacer una pregunta si es que se levanta la mano en la aplicación de Zoom.
- Se evaluará a los alumnos por tareas completas, se dejará material extra luego de cada sesión. Tareas prácticas y/o teóricas.

Normas de clase

- Todas las preguntas son bienvenidas, siempre que estén relacionadas al contenido de la clase en curso.
- No es bienvenida la burla o *bullying* entre alumnos y menos de parte de nuestros profesores.
- HackSpace es un ambiente de aprendizaje que busca fomentar el desarrollo **sin importar el sexo, raza, religión u orientación sexual**. Toda práctica de discriminación está prohibida dentro de este espacio.

Recomendaciones:

The image is a composite of two screenshots. The left screenshot shows a Zoom meeting interface. At the top, it says "Zoom Meeting ID: 277-019-445 Participant ID: 26". Below this, there are two video thumbnails: one for "Enzo Falcón" and another for "P. Benavides". A green box highlights the "P. Benavides" thumbnail, and a green banner below it says "You are viewing P. Benavides's screen". Below the video thumbnails, there is a toolbar with icons for "COVID-19.ipynb", "Archivo", "Editar", "Ver", "Insertar", "Entorno de ejecución", "Herramientas", and "Ayuda". The main area shows a code editor with the text `import pandas as pd`. The right screenshot shows a Google Colab notebook titled "COVID-19.ipynb" in a Mozilla Firefox browser. The browser's address bar shows <https://colab.research.google.com/>. The notebook's toolbar includes "File", "Edit", "View", "Insert", "Runtime", and "Tools". The main content area shows a table of COVID-19 data:

3	Confirmed	China	Western Pacific Region	278
4	Confirmed	Outside of China		5
5	Deaths	China	Western Pacific Region	NA

Below the table, there is a code cell with the following R code:

```
[ ] unique(data$WHO.region)

[ ] Western Pacific Region · South-East Asia Region · Region of the Americas ·
European Region · Eastern Mediterranean Region · African Region · Territories
► Levels:

▼ Agrupación de datos por región

[ ] df <- data[42:216, 2:ncol(data)]
df <- replace(df, is.na(df), 0)
df$X3.5.2020 <- as.integer(df$X3.5.2020)

# El operador '%>%' (pipeline) introduce como primer argumento a
# del lado izquierdo en la función del lado derecho.
group_data <- df[, 2:ncol(df)] %>%
  group_by(WHO.region) %>%
```