

CONFERENCIAS STATA LATAM 2022

Herramientas y aplicaciones  
estadísticas para Ciencia de Datos

# Metodología de datos sintéticos para modelos de *Machine Learning*

Franco A. Mansilla Ibáñez  
Septiembre 2022

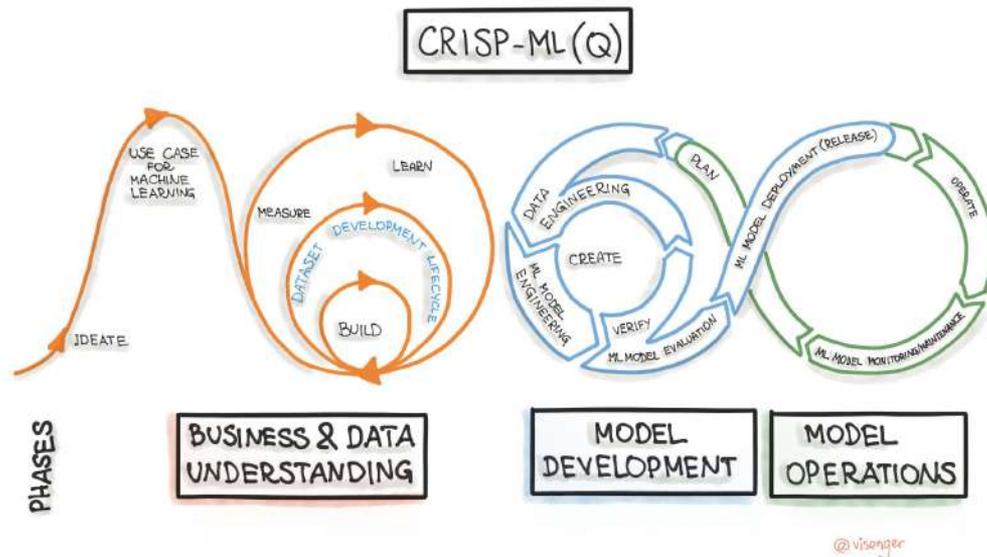


# Agenda

1. Introducción.
2. Pilares claves.
3. Algoritmos y Variables.
4. Datos sintéticos.
5. Aplicación en Stata 17.

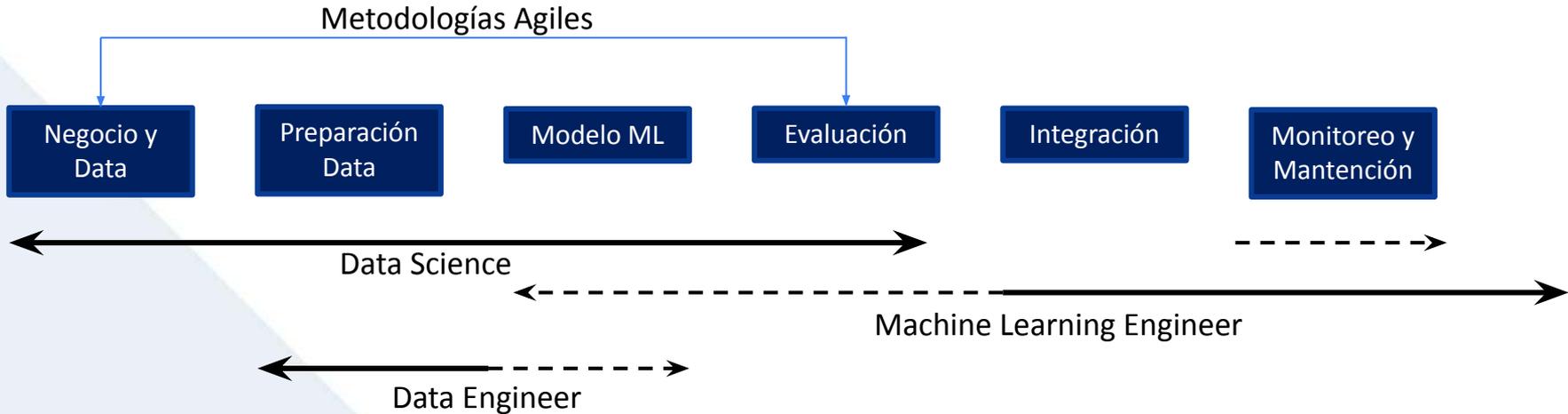
# Introducción

Figura 1: Ciclo de desarrollo de un proyecto de Machine Learning.



Fuente: [MLops](#)

## Pilares claves



—————> Habilidad Fuerte

-----> Habilidad Débil

# Algoritmos y Variables

Operacionales



Toma de Decisiones



Apoyo Gestión



Reducción de Tiempo

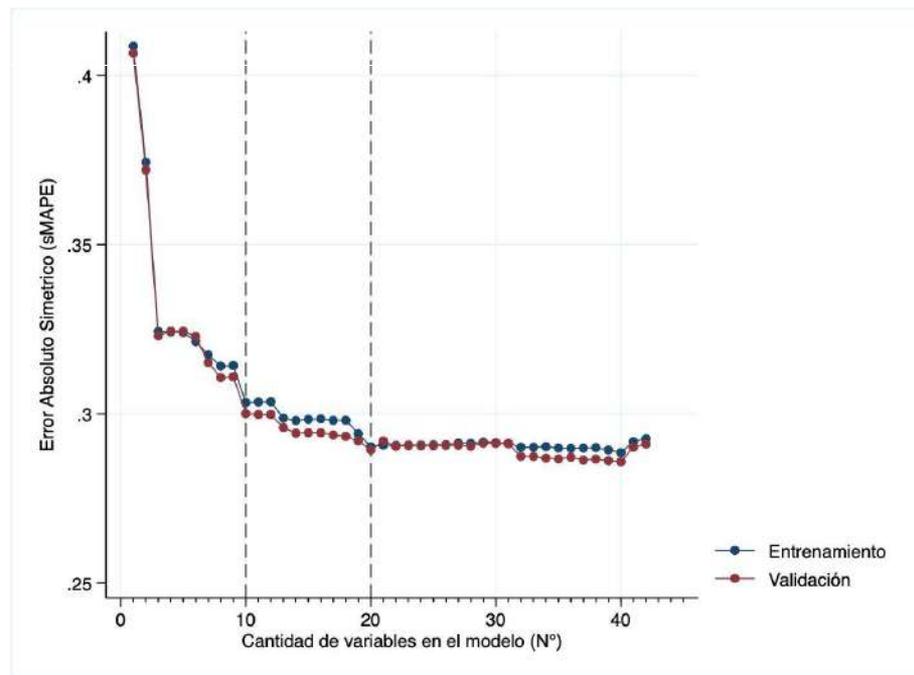


## Algoritmos y Variables (cont.)

Modelos Clásicos v/s Machine Learning

- BigData.
- Sesgo y Varianza.
- Capacidad Tecnológica.

Fuente: Elaboración propia.



## Datos Sintéticos

En la actualidad existe mucha investigación en metodologías para predecir datos en función a un contexto.

- Aleatoriedad.
- Aleatoriedad en función comportamiento (distribución de probabilidad).
- Anonimización y pseudoanonimización.
- Predicción por clúster.
- Predicción en imágenes

# Datos Sintéticos

¿para que se usan?

Nivelar Clases

Completar  
imágenes

Agregación de  
Datos

Predecir futuros  
comportamientos

Anonimizar  
datos

**Técnica:** Generative Adversarial Networks (GANs);

**Técnica:** Synthetic Minority Oversampling Technique (SMOTE)

# Aplicación

Código pSMOTE: <https://francomansilla.com/github>

The screenshot shows the GitHub repository page for 'fmansillaib/stata\_pSMOTE'. The repository is public and has 36 commits. The file list includes: 4. Imágenes, Código Ejemplo.do, Guía de Instalación y Uso (psmote)..., README.md, and psmote ado. The README.md file is selected, showing the title '[STATA]: proxy de SMOTE (pSMOTE)'. The description states it was created by Franco A. Mansilla Ibañez, Chile, and provides the website <https://www.francomansilla.com>. The version is 1.0-07/2022. The description lists two points: 1. The code psmote.ado will allow balancing classes of a dichotomous variable that is found unbalanced. 2. The code uses a clustering technique to define the method for creating synthetic samples as a function of those clusters.

## Do-file

```

1  *-----*
2  * CONFERENCIA DE STATA - SEPT. 2022 *
3  *-----*
4
5  *-----*
6  * Franco A. Mansilla Ibañez *
7  * www.francomansilla.com *
8  * www.software-shop.com *
9  *-----*
10 * Conferencia STATA 09/2022 *
11 *-----*
12
13 *-----*
14 * Definición pre-eliminar *
15 *-----*
16
17 clear all
18 set more off, permanently
19
20 * Cargar BD
21 import delimited "/Volumes/GoogleDrive-11186884732940162537/M1
22
23 * Renombrar variables
24 drop v1
25 ds *, varwidth(32)
26
27 global var_all = r(varlist)
28
29 local number=1
30 foreach i in $var_all {
31     rename `i' x`number'
32     local ++number
33 }
34
35 rename x3 fraude
36 drop x1 x2 x49 x50
37
38 *-----*
39 * Análisis de la Data *
40 *-----*
41
42 * 1. Tabulación de Fraude
    
```

# Conferencias Stata LATAM 2022

## Herramientas y aplicaciones estadísticas para Ciencia de Datos

Organiza:



Conozca más sobre STATA  
escaneando el código QR.

